The effects of adverse conditions on speech recognition by non-native listeners: Electrophysiological and behavioural evidence

Jieun Song

A thesis submitted in fulfilment of the requirements for the degree of *Doctor of Philosophy*

Department of Speech, Hearing and Phonetic Sciences

University College London

2018

Declaration

I, Jieun Song confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Jieun Song

Acknowledgements

First and foremost, I sincerely thank my primary supervisor, Paul Iverson, first for giving me this great opportunity to come to London to do this PhD, and providing incredible guidance and support throughout this journey. His office door was always open whenever I needed his help. I have learned so much from him in the past four years, and he has always been a great inspiration. I would also like to express my sincere gratitude to my secondary supervisor, Valerie Hazan. She was always there whenever I wanted to talk about my work and she has given me invaluable advice and support morally and academically. I am truly grateful.

I also would like to thank all staff members of SHaPS, including Jyrki Tuomainen for teaching me how to do EEG for the first time and giving me the opportunities to teach in his ERP course (not once, but three times!), Outi Tuomainen for her great advice on EEG testing and spontaneous speech, and Bronwen Evans for giving me the opportunities to brush up on my phonetic skills while teaching in her ear training sessions! I also thank Stuart Rosen for providing scripts to make speech-shaped noise and coming to our impromptu meetings about coherence. Great thanks also go to Steve Nevard, Andrew Clark, Dave Cushing and the IT team for their great technical help, and many thanks to Richard Jardine and the Finance team for their amazing admin support.

Special thanks go to my great friends and colleagues. I would like to thank Kathleen McCarthy and Emma Brint for their help with many things throughout the years (including being my best SSBE speakers!), and their moral support and great friendship. I feel so fortunate to have been able to work with you two! And of course,

my greatest thanks also go to my dear friends, Gisela Tomé Lourido, Faith Chiu and Petra Hödl. This rather tough journey was so much more fun and special because of them, and they have also been a great source of inspiration throughout my PhD. I also thank Sonia Granlund and Giulia Borghini for their great friendship and moral support. Thank you to all my lovely colleagues and friends in SHaPS, past and present: Rachel Zheng, Dan Kennedy-Higgins, Yue Zhang, Shiran Koifman, Wafaa Al-Shangiti, Mauricio Figueroa, Yasuaki Shinohara, Kurt Steinmetzger, Louise Stringer, Dong-Jin Shin, Albert Lee, Cristiane Hsu, José Joaquín Atria, Tim Schoof, Shego Wu, Anna Exenberger, and Laurianne Cabrera.

I also would like to thank my former supervisor, Ho-Young Lee at Seoul National University for his incredible support and encouragement throughout this process, and my other teachers at SNU Linguistics who taught me linguistics for the first time and sparked my interests in language.

My sincerest thanks of course go to my family – my parents, sister and brother – for their unconditional support, encouragement and belief. They always kept me going. Many thanks to my parents-in-law for being so proud of me. Finally, I am forever indebted to my husband for his unparalleled support throughout this tough journey, and for being so understanding when I was sacrificing so much of our time together. I would not have been able to do this without you.

Finally, my PhD was funded by the Kwanjeong Educational Foundation of South Korea and University College London. My sincere gratitude goes to them for making this thesis possible.

Abstract

This thesis investigated speech recognition by native (L1) and non-native (L2) listeners (i.e., native English and Korean speakers) in diverse adverse conditions using electroencephalography (EEG) and behavioural measures. Study 1 investigated speech recognition in noise for read and casually produced, spontaneous speech using behavioural measures. The results showed that the detrimental effect of casual speech was greater for L2 than L1 listeners, demonstrating real-life L2 speech recognition problems caused by casual speech. Intelligibility was also shown to decrease when the accents of the talker and listener did not match when listening to casual speech as well as read speech. Study 2 set out to develop EEG methods to measure L2 speech processing difficulties for natural, continuous speech. This study thus examined neural entrainment to the amplitude envelope of speech (i.e., slow amplitude fluctuations in speech) while subjects listened to their L1, L2 and a language that they did not understand. The results demonstrate that neural entrainment to the speech envelope is not modulated by whether or not listeners understand the language, opposite to previously reported positive relationships between speech entrainment and intelligibility. Study 3 investigated speech processing in a two-talker situation using measures of neural entrainment and N400, combined with a behavioural speech recognition task. L2 listeners had greater entrainment for target talkers than did L1 listeners, likely because their difficulty with L2 speech comprehension caused them to focus greater attention on the speech signal. L2 listeners also had a greater degree of lexical processing (i.e., larger N400) for highly predictable words than did native listeners, while native listeners had greater lexical processing when listening to foreign-accented speech. The results suggest that the increased listening effort

experienced by L2 listeners during speech recognition modulates their auditory and lexical processing.

Contents

List of Figures

List of Tables

Chapter 1 General Introduction

Understanding speech in a non-native language can be hard despite years of practice, especially without having sufficient experience speaking and hearing the language in a community where the language is spoken. When foreign language learners first come to target-language countries, communication with native speakers can feel quite daunting. The speech of native speakers sounds very different from the speech that they heard in the classroom; it feels too fast to follow and frequently deviates from the citation form (e.g., segment/syllable deletion). Non-native listeners commonly experience this difficulty even if they are able to recognise individual segments and words when they are carefully produced. Similarly, speakers with unfamiliar regional accents (e.g., Glaswegian accent) can be extremely difficult for non-native listeners to understand. In contrast, understanding less-fluent non-native speakers can be relatively easy especially if they share the same native language background.

To make matters worse, listening environments in real life are usually less than optimal. For example, non-native listeners struggle to understand speech in a noisy pub or even over the phone. As a result, they have to exert great effort to comprehend what the speaker is saying and frequently ask them to repeat what they have said. Non-native listeners can also feel mentally tired from straining to follow conversations, and cannot easily afford to perform another task while listening to speech, whereas such dual-tasking is more manageable for native listeners (e.g., driving while talking on the phone). Communication difficulties of non-native listeners can also lead to other problems in their daily life, such as failures to undertake tasks at work, social ineptitude, or feelings of loneliness.

Previous research has focused on understanding the interaction of first language (L1) and second language (L2) phonological systems to account for L2 speech learning difficulties, mostly using short fragments of careful "lab speech". Real-life factors affecting L2 speech recognition performance therefore remain largely unanswered. Specifically, can inaccurate knowledge of individual phonetic categories shown in laboratory conditions fully explain the difficulties experienced by L2 users in real life when listening to casual or continuous speech? One of the aims of this thesis is to examine L2 speech processing using more natural speech materials such as long, connected speech or spontaneously produced, casual speech. This thesis is also interested in exploring other real-life factors affecting speech intelligibility; background noise has been shown to be more detrimental to L2 listeners than to L1 listeners (e.g., see Lecumberri, Cooke, & Cutler, 2010 for a review), and in such degraded environments, listeners can benefit from listening to the accent that matches their own (e.g., Bent & Bradlow, 2003; Pinet, Iverson, & Huckvale, 2011; Wijngaarden et al., 2002). Exploring these effects can further our understanding of L2 speech recognition difficulties in everyday speech communication.

Furthermore, listeners can adapt their speech processing to fit the demands of the listening situation; for example, native listeners can draw on other sources of linguistic information or reduce the perceptual weight assigned to acoustic information in degraded listening conditions (e.g., Boothroyd & Nittrouer, 1988; McQueen & Huettig, 2012). Similarly, non-native listeners may adopt certain listening strategies to overcome their speech perception difficulties. For example, additional listening effort experienced by L2 listeners can affect their speech processing; it can interfere with

speech perception, or it may enhance their speech processing as a compensatory process that is not typically recruited for L1 processing (e.g., see Campbell & Sharma, 2013; Erb & Obleser, 2013; Peelle, Troiani, Grossman, & Wingfield, 2011 for listeners with hearing loss). That is, factors beyond the listener's linguistic knowledge can play an important role in L2 speech perception. The aim of this thesis is to investigate how increased speech comprehension difficulties of L2 listeners affect their speech processing especially in adverse listening conditions. Behavioural methods that are commonly used in L2 speech perception studies (e.g., phoneme identification/discrimination tasks, sentence recognition tasks) are more suitable for evaluating the outcome of speech perception/recognition processes (e.g., correct/incorrect). This thesis thus used electrophysiological methods which can measure the dynamics of speech processing at different levels (i.e., auditory, lexical). Issues on L2 speech perception in adverse conditions are discussed in detail in the following sections.

1.1 Second-language speech perception

Learning L2 speech sounds is difficult because individuals' early exposure to language alters neural organization such that it is specialised for that first language. That is, infants' perceptual representations become tuned into the phonological system of their native language, and thus become less sensitive to non-native sound contrasts during the second-half of the first year (e.g., Kuhl et al., 2006; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Werker & Tees, 1984). This loss of neural plasticity is thought to account for the age constraint on L2 speech learning; earlier is generally better. For example, L2 speakers who were first exposed to the target L2 at a later age

tend to speak the L2 with a stronger foreign accent (e.g., Flege, Munro, & MacKay, 1995; Flege, Yeni-Komshian, & Liu, 1999). The sensitive period for speech learning is thought to be the earliest compared to other linguistic domains (e.g., morphosyntax; Walsh & Diller, 1979).

It was long thought that listeners hear L2 sounds through a "sieve" of their L1 phonological system (Trubetzkoy, 1939), and L1-L2 interference is one of the core assumptions of L2 speech learning models. Best's Perceptual Assimilation Model (Best, 1995; Best, McRoberts, & Goodell, 2001) suggests that the difficulty of learning L2 sounds can be predicted by the articulatory similarity between L2 phonemes and existing L1 phonemes. For example, the discrimination of two L2 phoneme categories is expected to be easy if they belong to separate L1 phonemes ('Two-Category'), whereas it is expected to be poor if they are both good examples of the same L1 phoneme ('Single Category'). Flege's Speech Learning Model (James Emil Flege, 1995) is based on the assumption that L1 and L2 categories exist in the same phonological space. L2 categories are thus easier to learn if they are more dissimilar from the closest L1 categories and can thus fit in unoccupied regions in the phonological space (i.e., phonetic differences between the two sounds are better discerned), compared to similar or identical L2 sounds. Together, it is well established that L1 and L2 phonological systems interact and shape the ways we hear and produce L2 sounds.

However, L1-L2 interactions at the phonological level cannot fully account for L2 speech learning problems. For example, the difficulty in learning the English /r/-/l/

contrast experienced by native Japanese speakers is predicted by their representation of the third formant (F3) rather than their assimilation patterns (Hattori & Iverson, 2009). Native English listeners use F3 when distinguishing the phonemes, whereas Japanese listeners are more sensitive to F2 (i.e., a cue that is mostly irrelevant) because they have perceptual spaces that are not tuned for the English sound categories (Iverson et al., 2003). That is, it appears that L1 experience also interferes with L2 speech learning at a pre-linguistic, auditory level. Indeed, several neurophysiological studies have reported that the mismatch negativity response (MMN; i.e., automatic brain response to an odd stimulus in a repetitive sequence of identical stimuli) is larger for native than non-native phoneme contrasts (Brandmeyer, Desain, & McQueen, 2012; Dehaene-Lambertz, 1997; Näätänen et al., 1997; Winkler et al., 1999). Moreover, speakers of tone languages have more robust neural encoding of pitch even at the level of the brainstem compared to speakers of other languages (Krishnan, Gandour, Bidelman, & Swaminathan, 2009; Krishnan, Swaminathan, & Gandour, 2008; Krishnan, Xu, Gandour, & Cariani, 2005). However, results of this kind are not always found; some studies have found that cortical auditory-evoked potentials (e.g., N1 or the P1-N1-P2 complex) were not sensitive to whether or not the target sounds were phonologically distinctive in the listener's native language (e.g., Sharma, Marsh, & Dorman, 2000; Sharma & Dorman, 2000; Wagner, Shafer, Martin, & Steinschneider, 2013).

L2 speech recognition problems also exist beyond phoneme perception, because speech recognition involves understanding continuous, running speech, not words or phonemes in isolation. Listeners have to segment the incoming speech into isolated

words, using allophonic, phonotactic and prosodic cues marking word boundaries, as well as lexical and contextual information (e.g., Cutler & Norris, 1988; Mattys, White, & Melhorn, 2005; McQueen, 1998). L2 listeners can fail to locate the correct word boundaries because they cannot readily use those multiple sources of linguistic information, and may rely on their L1 segmentation strategies (e.g., Cutler, Mehler, Norris, & Segui, 1992). Word recognition can also be challenging for L2 listeners; correct lexical candidates may not be activated because the words are not in the listener's L2 lexicon. Furthermore, words from their L1 can be activated as well as L2 words (Spivey & Marian, 1999; Weber & Cutler, 2004), thereby increasing lexical retrieval effort (Schmidtke, 2014), and unnecessary lexical candidates can be activated due to inaccurate phoneme perception (e.g., Cutler, Weber, & Otake, 2006; see Lecumberri et al., 2010 for a review).

Furthermore, speech recognition processes are interactive such that problems at earlier processing stages (e.g., inaccurate phoneme perception, failure in lexical access) can be resolved by later semantic or syntactic processes (see Chapter 1.2 for details). However, these higher-level linguistic processes can also be less developed themselves in L2 listeners; native-like processing of complex grammatical structures is difficult to attain even for highly proficient L2 learners, whereas native-like lexical-semantic processing is relatively attainable (see Clahsen & Felser, 2006 for a review). Although the acquisition of syntactic and semantic processing skills itself is beyond the scope of speech perception research, inefficient semantic and syntactic processing can increase L2 speech recognition difficulties especially in suboptimal environments where sensory degradation needs to be compensated (see Chapter 1.3 for details).

1.2 Speech recognition in adverse conditions

In everyday life, speech communication often occurs in suboptimal listening environments such as noisy pubs or parties. Speech can be physically degraded by other signals in the background such as the speech from other talkers or environmental noises in shared spectro-temporal regions. This type of masking is called “energetic masking” (e.g., Brungart, 2001). Speech can also be distorted without any interfering sound sources - because of characteristics of the channel (e.g., filtering of telephone transmission) or reverberation (i.e., persistence of sound in an enclosed space). Adverse conditions can also originate in characteristics of speech production (referred to as ‘source degradation’ according to the taxonomy in Mattys, Davis, Bradlow, & Scott, 2012). For example, foreign-accented speech is difficult to understand particularly in noisy environments because it contains segmental and suprasegmental features that deviate from phonological representations of native speakers (e.g., Munro & Derwing, 1995; Munro, 1998). Previous research has also suggested that speech intelligibility in noise is determined by the interaction of the accents of the talker and listener (e.g., Bent & Bradlow, 2003; Pinet et al., 2011; Wijngaarden et al., 2002). That is, listeners generally find talkers who speak with the same accent as themselves easier to understand. Furthermore, casual speech that is produced in everyday conversations can be more difficult to understand than clear speech; casual speech contains more phonetic reduction phenomena (e.g., deletion, assimilation, and lenition) and has phonetic-acoustic characteristics such as faster speaking rate, smaller vowel dispersion that reduce intelligibility compared to clear speech (e.g., see Uchanski, 2008; Smiljanic & Bradlow, 2009 for reviews; see Mattys et al., 2012 for a general review of adverse conditions).

Speech comprehension can also be effortful when the listener's processing resources are taxed by concurrent tasks. Cognitive load is defined as "any load whose effect on speech recognition arises not from an energetic distortion of the signal, but from the recruitment of central processing resources due to concurrent attentional or mnemonic processing" (Mattys & Wiget, 2011). Cognitive load caused by a simultaneous task (e.g., visual search) can decrease native listeners' perceptual sensitivity to the acoustic signal as well as reduce their overall accuracy of speech perception (e.g., Mattys, Barden, & Samuel, 2014; Mattys, Brooks, & Cooke, 2009). More generally, listening in any adverse condition can be cognitively more demanding than listening in optimal environments, because listeners need explicit working memory (WM) related resources to resolve mismatches between their phonological representations and the acoustic input (e.g., distorted signals, accented speech; e.g., Rönnberg, Rudner, Lunner, & Zekveld, 2010). Furthermore, the presence of intelligible non-target signals (i.e., competing talkers) places additional demands on attention and cognitive control because the distracting speech needs to be ignored for the recognition of the target speech (i.e., informational masking; e.g., Cooke, Garcia Lecumberri, & Barker, 2008).

Nonetheless, native listeners are highly skilled at compensating for these difficulties to successfully decode the message of the speech signal. The speech recognition system of native listeners is robust, in that a loss of one source of information (e.g., distorted acoustic signals) can be overcome by relying on other sources of information that are available (i.e., lexical information, semantic context). For example, the "phoneme restoration" phenomenon (Warren, 1970) demonstrates that listeners use lexical knowledge when processing distorted speech; when a portion of an utterance

is replaced by a noise such as a cough, listeners report hearing the excised sound in the utterance along with the noise. Likewise, semantic and syntactic cues can help recognise ambiguous word forms or degraded speech sounds (e.g., Miller, Heise, & Lichten, 1951; Boothroyd & Nittrouer, 1988; Borsky, Tuller, & Shapiro, 1998; Connine, 1987; Kalikow, Stevens, & Elliott, 1977). That being said, there is disagreement among researchers as to how these findings are explained; by a direct top-down influence of higher-level information on lexical access, or late integration of different sources of information with a strict bottom-up information flow (e.g., McClelland & Elman, 1986; Norris, McQueen, & Cutler, 2000; Van Alphen and McQueen, 2001).

Moreover, native listeners can flexibly modulate their speech processing to fit the demands of the listening condition. For example, they can reduce the perceptual weight assigned to acoustic information during lexical competition if the acoustic signal is thought to be less reliable (McQueen & Huettig, 2012); when there were intermittent noise bursts in the speech signal, listeners were less certain about the word that they heard (i.e., looked less at onset-overlap and more at rhyme-overlap pictures in a visual-world eye-tracking experiment) even when the word was actually intact. Similarly, the relative weights of word segmentation cues can be assigned differently depending on the listening condition. Mattys et al. (2005) found that in optimal listening conditions, native listeners primarily replied on contextual and lexical cues to segment speech (e.g., a word was more likely to be segmented after another word than a non-word), whereas segmentation began to fall back on segmental cues (e.g., phonotactic or coarticulatory) when lexical information was impoverished. The

contribution of stress cues became strong only when both lexical and acoustic-phonetic cues were not reliable under severe noise.

1.3 L2 speech recognition in adverse conditions

Speech comprehension in noisy environments often feels doubly hard in a non-native language. This is because L2 listeners face "the dual challenges of *imperfect signal* and *imperfect knowledge*" (Lecumberri et al., 2010). Previous research has shown that the effect of adverse listening conditions is more detrimental to L2 listeners than to L1 listeners, even when L2 listeners are highly proficient bilingual speakers (e.g., Mayo, Florentine, & Buus, 1997; Nábělek & Donahue, 1984; Rogers, Lister, Febo, Besing, & Abrams, 2006; see Lecumberri et al., 2010 for a review). This normally occurs because L2 listeners' linguistic representations for the language are less developed compared to those of native listeners. Specifically, L2 listeners' phoneme perception in noise can be less accurate (e.g., Bradlow & Alexander, 2007; Hazan & Simpson, 2000) because their phonological representations can be less precise, or they may have not developed adaptive strategies to overcome the effect of noise at the segmental level (e.g., see Bradlow & Alexander, 2007 for a brief review). Moreover, this can occur because of L2 listeners' insufficient lexical, semantic and syntactic knowledge; non-native listeners are less able to benefit from higher-level linguistic cues to overcome acoustic degradation than are native listeners. For example, Mayo et al. (1997) found that highly proficient, late L2 listeners did not benefit from semantic-contextual cues (i.e., high predictability sentences) when listening to sentences in noise, whereas early bilingual and native listeners showed a strong benefit.

While the effect of additive noise on L2 speech recognition is well-known, there are other adverse conditions that have received little attention in the literature. In realistic communicative situations, L2 listeners often feel that speech communication is difficult even without any noise, because features of conversational speech such as faster speaking rate, segment/syllable deletion or assimilation can likewise be difficult for L2 listeners to overcome. That is, listeners need to be able to use acoustic-phonetic and higher-level linguistic cues (e.g., semantic context) to process such deviant word forms in casual speech (e.g., Ernestus, Baayen, & Schreuder, 2002; Gow, 2002). Because most previous research has been conducted using clear "lab speech" materials, little is known about the detrimental effect of casual speech on L2 speech perception.

In addition, the speech recognition system of L2 listeners may be more adversely affected by cognitive distractions (e.g., concurrent tasks or competing attention of distracting speech signals) than that of L1 listeners. It is largely because listening to L2 speech requires greater cognitive effort than listening to L1 speech (e.g., Schmidtke, 2014; see Indefrey, 2006; Stowe & Sabourin, 2005 for reviews), thereby depleting listeners' cognitive resources that could otherwise be available for dealing with the cognitive demands of the listening condition. Cognitive load could be expected to interfere with L2 speech recognition by reducing perceptual sensitivity to speech (e.g., Mattys & Palmer, 2015). However, listening effort can also be thought of as facilitating speech perception. For example, the engagement of some additional brain areas by listeners with hearing difficulties has been shown to improve speech comprehension (e.g., Erb & Obleser, 2013), and listeners can enhance their

representation of the acoustic signal through greater focused attention in competing-talker environments (e.g., Ding & Simon, 2012a). It is thus possible that increased listening effort and load experienced by L2 listeners may alter some aspects of speech processing, or result in the development of compensatory processes to help overcome their perceptual and comprehension difficulties.

1.4 The current thesis

The goal of this thesis is to investigate second-language speech recognition difficulties in adverse listening conditions. Various real-life factors were explored throughout this thesis, including casual speech style, speaker accent and background noise. This thesis was particularly interested in examining L2 speech perception difficulties that arise when listeners process more natural speech (i.e., continuous or casual speech). In addition to assessing L2 speech perception performance, the current thesis also investigated how L1 and L2 listeners modulate their auditory and lexical processing to overcome their recognition difficulties using EEG.

The first study of this thesis investigated the problems faced by non-native listeners when recognising casual speech. Chapter 2 details this behavioural speech-in-noise recognition experiment which assessed the speech recognition performance of native and non-native listeners for read and casually produced spontaneous speech. To measure the intelligibility of spontaneous speech, a new speech recognition task (a picture evaluation task) was developed using speech materials recorded via the DiapixUK task (Baker & Hazan, 2011). The aim of this study was to see if the detrimental effect of casual speech on speech recognition might be stronger for L2

listeners than for L1 listeners. Furthermore, listeners must understand casual speech in a range of native and non-native accents in everyday speech communication, and non-native listeners may show a greater advantage for their own non-native accent (e.g., Bent & Bradlow, 2003; Pinet et al., 2011) when native speakers talk casually. This study thus investigated how the effect of speech style interacts with the accents of the talker and listener to influence speech intelligibility.

While much of our knowledge of L2 speech perception has been based on studies using behavioural methods (e.g., phoneme identification or sentence recognition tasks), this thesis used electrophysiological methods as well as behavioural methods (Chapter 3 & Chapter 4) for several reasons. There is some evidence suggesting that L2 speech perception difficulties arise at an earlier, auditory level, than normally thought (Hattori & Iverson, 2009; Iverson et al., 2003; see Chapter 1.1 for details), and electroencephalography (EEG) can provide a means of measuring early auditory responses to speech. For example, the effect of native language experience has been found at auditory and sub-cortical levels of speech processing in neurophysiological studies (e.g., Näätänen et al., 1997; Krishnan et al. 2005). Furthermore, EEG was used in this thesis because new measures of cortical entrainment to speech provided a means to examine L2 speech processing at the auditory level for more natural speech (i.e., continuous speech) rather than focusing on one or two sounds as in typical ERP experiments. Specifically, a growing body of evidence suggests that when listeners process continuous speech, oscillations in the auditory cortex become phase-locked (i.e., entrained) to slow amplitude modulations in the speech signal in delta (1-3Hz) and theta (4-8Hz) frequency ranges (e.g., Ahissar et al., 2001; Luo & Poeppel, 2007;

Peelle, Gross, & Davis, 2013). This neural tracking of the speech envelope has been observed in single-trial neural recordings using continuous speech stimuli such as stories (e.g., Ding & Simon, 2012a; Howard & Poeppel, 2010; O'Sullivan et al., 2015).

Moreover, EEG can measure the dynamics of speech processing. That is, EEG measures can tap into speech recognition processes at different stages as they occur. This can be useful for the purpose of the present work, because such measures can reveal online processing mechanisms that listeners use to overcome their speech recognition difficulties. Specifically, the N400 component of the event-related brain potential (Kutas & Hillyard, 1980) was used to measure neural effort that listeners exert for lexical processing in given semantic contexts, which is difficult to measure with behavioural speech recognition tasks which only assess the outcome of speech recognition processes.

Chapter 3 describes the first EEG study that examined neural entrainment to the amplitude envelope of speech while subjects listened to their native language, second language or a language that they did not understand (i.e., native English and Korean subjects listening to English, Korean and Spanish). This study originally set out to develop EEG methods that measure how second-language listeners process continuous speech, because neural entrainment to speech was expected to be sensitive to the listener's native language experience; the response has been thought to be related to syllable-parsing or speech comprehension (e.g., Giraud & Poeppel, 2012; Peelle et al., 2013). Using this cross-linguistic design, this study was also able to investigate the much-debated issue on the link between cortical entrainment to speech and speech

intelligibility (e.g., Peelle et al., 2013; Howard and Poeppel, 2010; see Chapter 3.1.2 for details) without altering the acoustic properties of the speech signals. This EEG measure was also used in the following EEG study, but in a different listening environment.

Chapter 4 describes the second EEG study which focused on investigating how listeners cope with their speech recognition difficulties. Specifically, this study examined speech processing by L1 and L2 listeners in a competing-talker environment (i.e., two talkers were presented to separate ears) in which the target and distracting speakers had an L1 or L2 accent. Speech recognition in this environment is more effortful for L2 listeners because of the informational masking caused by the distracting talker, combined with their intrinsic L2 speech recognition problems. This study used measures of neural entrainment and N400 as well as a behavioural speech recognition task (i.e., detection of anomalous sentences) to more comprehensively examine how L1 and L2 listeners modulate their processing in difficult listening conditions at auditory and lexical levels.

Chapter 2 The effect of casual speech for non-native listeners[1]

2.1 Introduction

Non-native listeners commonly experience increased speech comprehension difficulties when speakers talk casually, although they can be highly accurate at comprehending clear, read sentences in a language exam. This is mostly because speech used in casual conversations contains a number of phonetic variations such as deletion, assimilation, and liaison (e.g., Johnson, 2004), which can impede speech recognition by non-native listeners who are normally less able to compensate for such casual speech processes (e.g., Tuinman, Mitterer, & Cutler, 2011). Although this problem seems obvious, the effect of casual speech on L2 speech perception is not well understood. Furthermore, listeners often encounter casual speech in a range of native and non-native accents. Listeners tend to find those who speak with the same accent as themselves more intelligible than others under noise (e.g., Bent & Bradlow, 2003; Pinet et al., 2011), and because native speech tends to be more reduced when produced casually, speech communication between non-native speakers can sometimes feel less effortful especially when they share the same L1. This indicates that the problem of understanding casual speech in an L2 may also depend on the accents of the talker and listener.

This chapter describes a behavioural speech-in-noise recognition experiment which assessed the speech recognition performance of native English (L1) and Korean (L2)

[1] Part of this work has been published in a preliminary form in the proceedings of the 18th International Congress of Phonetic Sciences (Glasgow, UK) as: Song, J., and Iverson, P. (2015). Measuring speech-in-noise intelligibility for spontaneous speech: the effect of native and non-native accents.

listeners for read and casually produced spontaneous speech. This study used new methods to measure speech recognition performance for spontaneous speech as well as read speech (a picture evaluation task). The aims of this study were to see if the detrimental effect of casual speech could be stronger for L2 listeners than for L1 listeners, and to investigate how the recognition of read and casual speech is affected by the accents of the talker and listener. Listeners are expected to show an intelligibility advantage for their own accent when listening to casual speech as well as read speech, but non-native listeners might display a stronger advantage for their own non-native accent for casual speech. Furthermore, the listener's familiarity with the talker's accent (e.g., Adank, Evans, Stuart-Smith, & Scott, 2009) and the acoustic-phonetic similarity between the talker's and listener's accents (e.g., Pinet et al., 2011) have been proposed to account for accent intelligibility. The current study also examined how these factors contribute to intelligibility differences between accents.

2.1.1 The recognition of casual speech

2.1.1.1 The recognition of reduced word forms

The speech that we hear in everyday life is far more variable than the carefully read speech that is elicited in a laboratory setting. Specifically, phonetic reduction processes such as deletion, assimilation, or lenition (e.g., /t, d/ deletion as in 'just' [ˈdʒʌs] and 'next' [ˈneks]; schwa deletion as in 'summary' [ˈsʌmɹi] and 'personal' [ˈpɜːsnəl] commonly occur in natural speech. The occurrence of these processes can be conditioned by linguistic factors such as phonological or morphological context; for example, schwa deletion occurs when the resulting onset consonant cluster forms a

sonority rise[2] (e.g., Hooper, 1978). However, it is also affected by extra-linguistic factors such as speaking rate and word frequency[3], with more reduction occurring in faster speech and more frequent words (e.g., Guy, 1980; Fosler-Lussier & Morgan, 1999; Jurafsky, Bell, Gregory, & Raymond, 2000).

Reduction phenomena are also more likely to occur in spontaneous casual speech when speakers exert less articulatory effort. Conversational speech involves more extreme cases of reduction in which multiple phonemes or syllables are deleted. For example, the word 'particular' can even be pronounced as [pʰtʰɪkɚ] in spontaneous speech (Johnson, 2004). Johnson (2004) reported that in a corpus of American English conversational speech more than 60 % of words deviated from their citation forms by at least one phone, and 28 % of words deviated by two or more phones. Phonetic reduction can be explained by the 'hyper- and hypo-articulation theory' (H&H theory, Lindblom, 1990) which suggests that speakers can change their production along the continuum of hyper- (i.e., clear speech) and hypo-articulation (i.e., reduced speech) under two competing constraints – minimum articulatory effort and perceptual saliency. That is, speakers can produce reduced forms by speaking casually as long as they can be readily understood by the listener, but they can speak more clearly when it is necessary to accommodate the needs of the listener (e.g., a hearing-impaired listener).

[2] Speech sounds can be ranked according to their sonority which is roughly correlated with loudness. In the above-mentioned examples of schwa deletion, the second member of the resulting consonant cluster is higher in sonority than the first member (i.e., liquids /ɹ/ > nasals /m, n/ > obstruents /s/; Clements, 1990).

[3] There are a variety of extra-linguistic factors affecting the occurrence of phonological variation such as the gender or age of the speaker.

Despite the large amount of variation in speech, listeners are highly skilled at processing such deviant word forms during speech recognition. Models of spoken-word recognition make different proposals regarding how pronunciation variants are recognised. Abstractionist models (e.g., Marslen-Wilson, 1987; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Norris, 1994) argue that only canonical word forms are stored in the mental lexicon. When hearing a reduced/deleted word form, listeners therefore need to reconstruct the intended word form from the variant. Previous research has suggested that listeners are able to recognise massively reduced word forms if semantic/syntactic context is available (e.g., in a context of several words; Ernestus et al., 2002; Kemps, Ernestus, Schreuder, & Baayen, 2004). The interactive model of spoken-word recognition, TRACE (McClelland & Elman, 1986), explains top-down influence by allowing higher-level linguistic information to exert effects on pre-lexical processing during lexical access[4]. Furthermore, listeners can use fine phonetic detail to recognise a reduced form. For example, Gow (2002) found that listeners were sensitive to subtle phonetic differences between an assimilated, underlyingly coronal segment [p] in '*righ**t*** [ɹaɪp] *berries'* (i.e., place assimilation before /b/), and a noncoronal segment [p] in '*ri**p**e* [ɹaɪp] *berries*', thereby resolving potential lexical ambiguity caused by the place assimilation[5].

[4] In contrast to TRACE, autonomous abstractionist models such as Shortlist (Norris, 1994; Norris, McQueen, Cutler, & Butterfield, 1997) argue that best lexical candidates are selected based on the degree of fit between the acoustic input and lexical candidates without lexical feedback. However, these abstractionist models agree that spoken-word recognition is performed based on competition among multiple lexical candidates that are simultaneously activated.

[5] Lahiri and Marslen-Wilson (1991, 1992) proposed that in the mental lexicon, words are represented with values of some features of sounds not specified. For example, the feature 'coronal' is underspecified in the above example of place assimilation. The word 'right' as well as 'ripe' can therefore be activated upon hearing the assimilated variant [ɹaɪp].

In contrast, episodic models (e.g., Bybee, 2001; Goldinger, 1998; Hawkins, 2003; Johnson, 1997; Pierrehumbert, 2001) suggest that all pronunciation variants of each word are stored in the mental lexicon. These multiple 'exemplars' originate from all variants that the listener has encountered in the past. In these models, a reconstruction process is not necessary to recognise deviant word forms as all exemplars (i.e., variants) that are relevant to the incoming acoustic input are activated. Although the evaluation of these models is beyond the scope of the present study, it is widely acknowledged that listeners are proficient in dealing with pronunciation deviants during speech perception.

2.1.1.2 Intelligibility of clear speech versus conversational/casual speech

Aside from having more phonetic reductions than clear speech, casual/conversational speech has other global acoustic characteristics. Previous studies have mostly been interested in finding the acoustic-phonetic features of clear speech compared to casual/conversational speech, rather than focusing on conversational speech alone (e.g., see Uchanski, 2008; Smiljanic & Bradlow, 2009 for reviews). In this line of research, clear speech is elicited by asking talkers to read speech materials (e.g., sentences) following specific instructions; as if they are talking to non-native or hearing-impaired listeners or in the presence of background noise. In contrast, conversational speech is recorded by asking talkers to read the same kind of materials in a causal speaking style – as if they are talking to a friend.

Previous studies have shown that compared to casual speech, clear speech is produced with a slower speaking rate, more frequent and longer pauses, increased average pitch

and pitch range, greater overall intensity, greater energy at high frequencies (above 1000 Hz) of long-term spectra, and higher peaks in the 1-3 Hz range of modulation spectra (see Uchanski, 2008; Smiljanic & Bradlow, 2009 for reviews). Vowel space expansion has also been found in clear speech (e.g., Ferguson & Kewley-Port, 2002, 2007; Moon & Lindblom, 1994; Picheny, Durlach, & Braida, 1986). That is, speakers increase the distance between vowel categories in the vowel space to make them perceptually more distinct from one another. Changes in consonantal properties have also been reported for clear speech such as longer voice onset time (VOT) values in unvoiced stops or increased consonant-to-vowel relative power ratios (e.g., Bradlow, Kraus, & Hayes, 2003; Krause & Braida, 2004; Picheny et al., 1986).

Previous research has also found that most of these acoustic-phonetic properties of clear speech indeed enhance speech intelligibility (e.g., Hazan & Markham, 2004; Liu & Zeng, 2006; see Smiljanic & Bradlow, 2009 for a review). Overall, both normal-hearing listeners in the presence of background noise and hearing-impaired listeners have been shown to benefit from clear speech compared to conversational speech for various materials (e.g., syllables and sentences). The average clear speech intelligibility gain for normal-hearing and hearing-impaired listeners was 20 and 26 percentage points, respectively, in Payton, Uchanski, and Braida (1994). Conversely, one could expect that casual speech can reduce speech intelligibility compared to clear speech or read speech particularly in adverse conditions (e.g., normal-hearing listeners in noise and hearing-impaired or non-native listeners), although the intelligibility of conversational speech itself has not been an area of focus in previous research. That

is, conversational speech can be seen as one form of adverse conditions (Mattys et al., 2012).

2.1.1.3 The processing of casual speech

That being said, casual speech is an ordinary style of speech that listeners encounter in everyday life, and thus it does not seem very obvious that listeners would have difficulty understanding casual speech, especially without any noise. This is because native listeners can cope with reduced phonetic information in casual speech, due to the redundancy of the speech signal; they draw on a variety of cues – sub-lexical cues (e.g., acoustic-phonetic), phonological context and higher-level linguistic information (e.g., lexical or semantic-contextual) – to decode the structure and meaning of speech (e.g., Ernestus et al., 2002; Gaskell & Marslen-Wilson, 1996; Gow, 2002; McClelland & Elman, 1986). A clear speech advantage over casual speech has only been found in adverse listening conditions for normal-hearing listeners (e.g., see Uchanski, 2008 for a review).

Furthermore, listeners can adopt speech recognition strategies that are optimal for processing casual speech. Specifically, an eye-tracking study by Brouwer, Mitterer, and Huettig (2012) found that when the speech signal contained great phonetic reduction overall, listeners penalised acoustically non-matching lexical competitors less strongly. Specifically, listeners normally fixated a canonical form competitor (e.g., *benadelen* for a Dutch word *beneden* “downwards”) more than a reduced form competitor (e.g., *meneer* for the reduced form [məneːə]) when they only heard canonical forms during the experiment, as expected by the well-established

phenomenon where lexical candidates with initial overlap are more strongly activated than those with medial or final overlap (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). However, when reduced word forms were intermixed with canonical word forms in the experiment, listeners fixated a reduced form competitor as much as a canonical form competitor, regardless of whether or not the target word was actually reduced.

2.1.2 The use of spontaneous speech for perception studies

It is important to note that the casual speech elicited in the aforementioned clear speech literature differs from spontaneously produced, casual speech because speakers were simply asked to read linguistic materials 'in a casual speaking style' in those studies. This does not appear to be an ideal way of obtaining naturally produced, casual speech because it is still read speech, and thus it is less likely to contain features of reduced speech that are found in realistic communicative settings. However, the advantage of using this method is that one can have complete control over what speakers say. It is thus possible to elicit target words/phonemes while controlling for factors which can influence the acoustic-phonetic realisations of the target items such as phonological and prosodic context or lexical features (e.g., frequency, first/second mention).

In contrast, spontaneous speech is not read (i.e., it is unscripted), and is often elicited while a speaker is conversing with other interlocutors in naturalistic communicative settings. Casual speech that was used in the current study was spontaneous speech that was elicited between normal-hearing talkers/listeners in such natural situations, which likely has more features of casual speech such as segment or syllable deletion, faster

speaking rate, and reduced vowel space compared to casual read speech used in the previous clear speech literature.

Various methods have previously been used to obtain more natural speech that is used in everyday situations (see Warner, 2012; Baker & Hazan, 2011 for reviews). For example, the Buckeye Corpus (Pitt et al. 2007) is a conversational speech corpus of American English, which was obtained through interviews. Similar corpora also exist for other languages (e.g., the Spoken Dutch Corpus, Oostdijk, 2000; the Seoul Corpus, Yun et al., 2015). The Switchboard corpus (Godfrey, Holliman, & McDaniel, 1992) is a collection of telephone conversations between speakers of American English from around the U.S. Each of these methods has its own advantages and disadvantages in terms of the degree of naturalness and recording quality, but conversational speech from such corpora likely contains more features of casual speech compared to casual read speech. However, one caveat is that the speech from such corpora is completely unscripted and highly variable. It can thus be more difficult to find productions of target words across different speakers that meet specific criteria for a study (e.g., phonological and prosodic context). That is, a dilemma between naturalness of speech and having control over what speakers say persists.

In an attempt to have more control over what speakers say, speakers can be asked to retell a story that they have read using their own words (e.g., IViE Corpus; Grabe, Post & Nolan, 2001; The CHAINS corpus; Cummins, Grimaldi, Leonard, & Simko, 2006). This can elicit spontaneous casual speech which still contains target keywords. Another way of getting around the dilemma could be having speakers read written

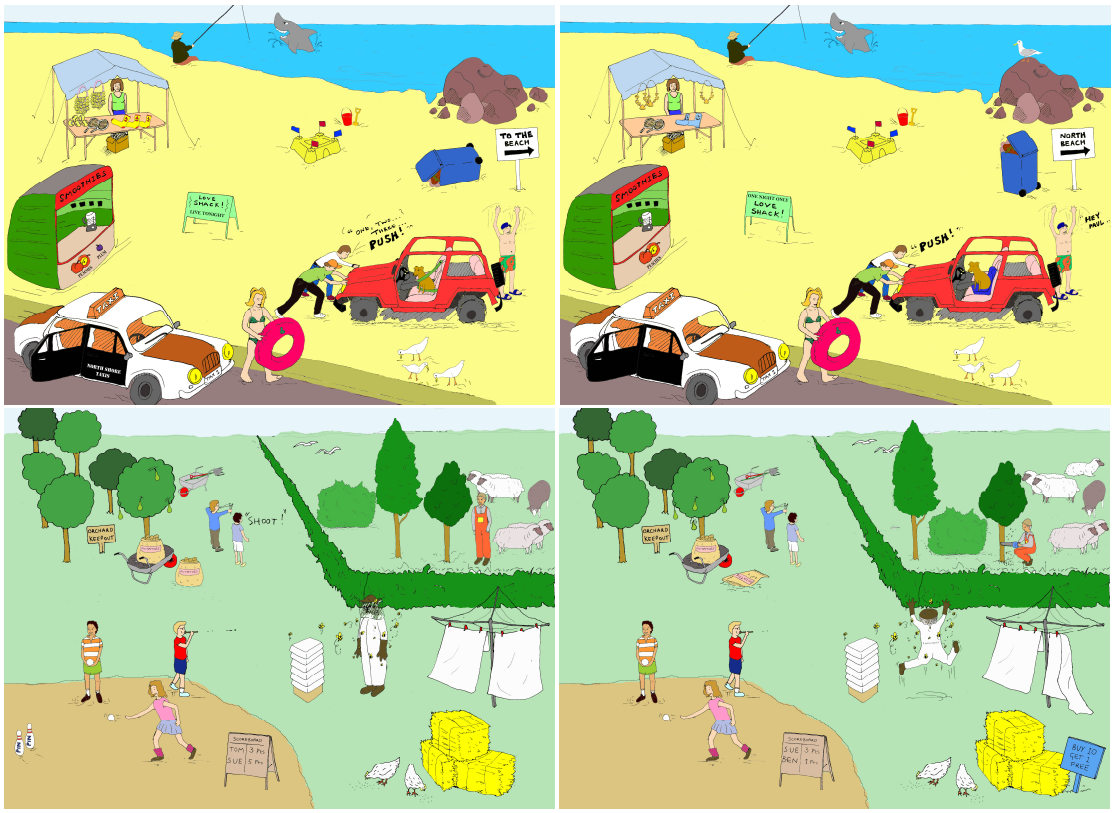
transcriptions of the utterances they had spontaneously produced (Mehta & Cutler, 1988; Haynes, White, & Mattys, 2015). This elicitation method can create a complete match between read and spontaneous speech, but it still involves finding and selecting appropriate utterances/sentences from the spontaneous speech produced by each speaker.

Another method of obtaining casual but relatively controlled speech is to record speakers while they perform a problem-solving task together. The advantage of using this method is that it can elicit spontaneous interactions between two speakers while having some control over the lexical content of the speech. Sudoku puzzles or crosswords have been used (Cooke & Lu, 2010; Crawford, Brown, Cooke, & Green, 1994), but these tasks involve relatively limited interactions between speakers (Baker & Hazan, 2011). The Map Task (Anderson et al., 1991) has also been widely used; spontaneous dialogues between two speakers are recorded while one speaker gives directions on a map to the other speaker. Because the two maps that each speaker has are slightly different, they likely converse interactively asking for clarification. In addition, the maps are designed to elicit productions of certain words/phonemes.

More recently, the Diapix Task (Van Engen et al., 2010) was developed to elicit spontaneous interactive speech between two interlocutors. In this task, each speaker is given a different version of the same cartoon-style picture and they have to find differences between these two pictures together without looking at what is in each other's picture (also called a 'spot the difference' task). The pictures were designed to induce the production of certain keywords (i.e., minimal pairs). Meanwhile, speakers

can converse with each other freely in a realistic communicative setting, without one speaker having to take a more leading role than the other as in the Map Task. Furthermore, the richness and complexity of the pictures make speakers produce a greater variety of utterance types such as declarative sentences, questions, answers and exclamations (Van Engen et al., 2010), in contrast to the aforementioned problem-solving tasks which normally elicit limited types of sentence structures such as requests and demands (e.g., the Map Task). However, Diapix still allows for maintaining some control over the lexical content of the discourse, similar to the previous methods.

Figure 2-1: DiapixUK picture materials: pairs of beach scene 3 (top) and farm scene 4 (bottom)



The present study used spontaneous speech recorded via the DiapixUK task (Baker & Hazan, 2011). The DiapixUK task contains a larger set of pictures (12 pairs; examples shown in Figure 2-1) than does the original version of the Diapix (Van Engen et al., 2010), and the pictures can also be easily adapted to one's own research purposes. The current study used casual spontaneous speech that was elicited between normal-hearing talkers in a 'no-barrier' listening condition (i.e., without any background noise or signal degradation; see Chapter 2.2.2. for further details). Compared to other challenging communicative settings used with the DiapixUK task (e.g., one speaker hearing the other through a vocoder or with babble noise; e.g., Hazan & Baker, 2011), talkers are expected to speak more casually in this condition.

Spontaneous speech has been more extensively used to investigate aspects of speech production such as clear speech modifications in response to the needs of the listener (e.g., hearing-impaired listeners or listeners in adverse listening conditions; e.g., Hazan & Baker, 2011; Hazan, Grynpas, & Baker, 2012; Tuomainen, Hazan, & Romeo, 2016), but it is generally more difficult to use spontaneous speech for perception experiments. Although words or phrases extracted from spontaneous speech data have been previously used for perception experiments (e.g., Ernestus et al., 2002; Hazan et al., 2012; White, Mattys, & Wiget, 2012), very few studies have used longer stretches of speech such as sentences or utterances from spontaneous speech, including Haynes et al. (2015) that used spontaneous utterances in a speaking style judgment task. This is likely because extracting appropriate sentences from uncontrolled spontaneous speech is more difficult, whereas extracting phonemes or words is relatively more feasible using those tasks that are designed to elicit target keywords.

The present study developed an effective way to evaluate speech recognition performance for utterances from spontaneous speech; instead of asking listeners to repeat back what they had heard as in typical speech recognition tests that use read speech materials, listeners were presented with a picture on the screen as they listened to a stimulus and they then had to decide whether or not what they heard matched the picture. This paradigm is particularly appropriate because spontaneous speech does not lend itself very well to word-by-word repetition due to its unstructured and dynamic nature. For example, the number of content words and syntactic complexity can vary between stimuli. Meanwhile, using the DiapixUK task allowed for maintaining some control over the lexical and semantic content of speech across speakers.

2.1.3 The recognition of spontaneous casual speech by non-native listeners

Importantly, the present study was interested in investigating how speech recognition by second-language listeners is affected by speech style (i.e., spontaneous casual speech vs. read speech). Although L2 listeners can feel that casual speech is far more difficult to understand than the speech that they hear in the classroom, most of our knowledge of L2 speech perception is based on investigations using clear read speech. One could expect that features of casual speech such as deletion, assimilation, liaison or less salient acoustic-phonetic cues are more difficult for L2 listeners to overcome than for L1 listeners. As discussed above, native listeners are able to exploit a variety of linguistic knowledge to recognise reduced forms – low-level phonetic detail, phonological context, and lexical-semantic cues. In contrast, there is ample evidence showing that L2 listeners are more adversely affected by difficult listening conditions

such as background noise because of their insufficient/inaccurate perceptual and linguistic representations at all these levels (see Chapter 1.3 for details).

Although the detrimental effect of casual speech as a whole is not known for non-native listeners (see Bradlow & Bent, 2002 for the effect of clear speech for non-native listeners), a small number of studies have investigated the perception of specific casual speech processes by L2 listeners. For example, Tuinman et al. (2011) found that Dutch learners of English had difficulty processing /r/-insertion in British English (e.g., 'idea(r) of' [aɪdɪəɹ ɒv]) using the relevant acoustic cue (i.e., duration) as this process is absent in Dutch, whereas German learners of Dutch were able to process /t/-deletion in Dutch similarly to native Dutch listeners because the same process exists in German (Mitterer & Tuinman, 2012). However, their perception deviated from that of native Dutch listeners when /t/-deletion occurred in verbs, where this process does not apply in German. This demonstrates that L1 interference extends to the domain of casual speech processes (also see Darcy, Ramus, Christophe, Kinzler, & Dupoux, 2009).

2.1.4 Talker-listener accent interactions for casual speech

As mentioned previously, speech recognition is also affected by the accents of the talkers and listeners especially in noisy conditions. That is, listeners understand talkers who speak with the same accent as themselves more easily than others (e.g., Bent & Bradlow, 2003; van Wijngaarden et al., 2002; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002; Imai, Walley, & Flege, 2005; Pinet et al., 2011). This accent effect can arise because L2 listeners and talkers from the same L1 background share an interlanguage (Bent & Bradlow, 2003). Specifically, they share extensive linguistic

knowledge with each other which includes a variety of phonetic and phonological aspects of both the (incomplete) L2 and L1, whereas the shared linguistic knowledge between native and non-native talkers/listeners only covers part of the L2 knowledge (i.e., L2 knowledge that non-native listeners have acquired). A non-native listener with a shared L1 is therefore thought more likely to identify the pronunciation of a non-native talker (e.g., sounds that were produced as similar to their L1 phoneme categories) as the talker intended than a native listener (Bent & Bradlow, 2003). Other work on talker-listener accent interactions has suggested that accent intelligibility is determined by acoustic-phonetic similarity between talker and listener accents (Pinet et al., 2011). That is, listeners can find accents that are more acoustically similar to their own accent easier to understand than those that are more distant. For example, English-French bilinguals and French experienced learners of English whose accents were more similar to the accent of native English speakers did not show a clear intelligibility advantage for French-accented English (Pinet et al., 2011), despite sharing their L1 phonology as well as part of their L2 phonology with the talkers. This suggests that the amount of L2 experience also influences talker-listener accent interactions.

Alternatively, this accent benefit could be explained by the listener's familiarity with the talker's accent. For example, Adank et al. (2009) found that Glaswegian listeners, who were familiar with Standard Southern British English (SSBE) through the media and interactions with Southerners were able to comprehend SSBE and Glaswegian English equally fast in background noise, whereas Southern British listeners were slower with the less-familiar Glaswegian accent than with SSBE. Previous research has shown that experience with a particular accent enables listeners to learn to map

variant forms of that accent into their underlying forms (e.g., Sumner & Samuel, 2009). Similarly, listeners have been shown to adapt their perceptual system to a novel accent even after brief exposure to that accent (e.g., Clarke & Garrett, 2004).

It is important to note that all these previous findings were based on intelligibility of read speech materials. However, one could expect that these talker-listener accent interactions found in read speech will extend to spontaneous casual speech. Characteristics of a talker's accent should be abundant in spontaneous casual speech because it could reflect L1 interferences not only in segmental and suprasegmental aspects, but also in connected speech processes. Casual speech processes that are transferred from the speaker's L1 (e.g., Spanish speakers can apply the lenition process of Spanish to English; Flege & Davidian, 1984) might be more easily processed by listeners from the same L1 background, contributing to the accent advantage. On the contrary, non-native listeners have been shown to have difficulty compensating for casual speech processes produced by native speakers, unless they have the same processes in their native language (e.g., Tuinman et al., 2011).

Moreover, due to the characteristics of casual speech produced by non-native listeners, a greater talker-listener accent interaction may arise for casual speech. Bent & Bradlow (2003) found an intelligibility benefit between non-native talkers and listeners even when they did not share an L1. For example, native Chinese listeners found native Korean talkers more intelligible than native English talkers. The authors suggested that this occurred because non-native speakers are typically less likely to apply reduction processes such as deletion or failure to release word-final stops than are native

speakers. That is, more salient phonetic cues in non-native speech can be beneficial for non-native listeners, who also have not learned such casual speech processes. Casual speech could then provide an interesting testing ground for further investigating the talker-listener accent interaction phenomenon.

2.1.5 Aims of the current study

The aim of the current study was to investigate how speech recognition by native and non-native listeners is modulated by speech style and the accents of the talker and listener. It is expected that the effect of casual speech is more detrimental to non-native listeners because they are not able to compensate for degraded or variable phonetic information in casual speech as freely as native listeners. One could also expect that listeners understand their own accent better than others' when listening to casual speech as well as read speech, but it is possible that this accent effect is stronger for casual speech; additional phonetic and phonological features in casual speech such as connected speech processes might add to the advantage. Moreover, non-native listeners may display a greater intelligibility benefit for their own non-native accent when listening to casual speech, because native-accented speech may contain more reduced pronunciation variants than non-native speech when it is produced casually.

In this study, native English and Korean subjects listened to read sentences and spontaneous utterances in noise, in a native English accent (Standard Southern British English) and two non-native English accents (Finnish and Korean-accented English). The Korean subjects (i.e., listeners) were L2 learners of English with a similar amount of English experience (i.e., living in English-speaking countries for an average of 9.8

months; see Chapter 2.2.1). Background noise was added to the stimuli to avoid ceiling effects. Speech recognition performance was measured with the picture evaluation task for both read and spontaneous speech materials.

Moreover, the present study explored how the similarity between talkers' and listeners' accents determines accent intelligibility; acoustic analyses were conducted on the sentence recordings of the Korean subjects and three groups of talkers using the ACCDIST metric (Huckvale, 2004, 2007a,b) to measure the acoustic-phonetic similarity between the accents. Finnish-accented English was added as one of the talker accents to examine a wider range of accent familiarity and talker-listener accent similarity. Specifically, Finnish-accented English was unfamiliar to the Korean listeners, whereas SSBE and Korean-accented English were both familiar to them. English listeners were only familiar with their own accent (i.e., SSBE). Finnish-accented English was also chosen among other non-native accents of English that are unfamiliar to Koreans, because it was expected to be substantially different from Korean-accented English in terms of their phonetic and phonological characteristics; Korean and Finnish do not have great phonetic similarity in that they do not have any genetic relationship and have widely different segmental and suprasegmental properties (e.g., Suomi, Toivanen, & Ylitalo, 2008).

2.2 Methods

2.2.1 Subjects

Nineteen monolingual native speakers of Standard Southern British English (mean age: 22.5 years, age range: 18-28 years) and 24 monolingual native speakers of Korean

(mean age: 28 years, age range: 19-40 years) participated in the experiment. All the participants had no self-reported hearing or language disorders and were living in London at the time of testing. The Korean subjects reported that they had started learning English at school in South Korea from the age of 11 years old on average (range: 5-14 years) and they had been living in English-speaking countries (i.e., mostly England[6]) for an average of 9.8 months (range: 1-36 months) as adults. None of them had resided in English-speaking countries before becoming adults.

2.2.2 Stimuli and apparatus

Apart from the subjects (i.e., listeners) who participated in the perception experiment, two female speakers of each of these three accents – Standard Southern British English, Finnish-accented English and Korean-accented English – took part in the recording (age range: 18-30 years, mean age: 24.7 years). The British speakers were monolingual native speakers of Standard Southern British English. The Finnish speakers reported that they had never lived in English-speaking countries, but they had been learning English in Finland since they were nine years old. The Korean speakers reported that they had been living in London for approximately eight to twelve months and had started learning English at school in South Korea when they were twelve. The speakers in each pair were close colleagues, friends or sisters.

In order to obtain spontaneous speech, the DiapixUK task (Baker & Hazan, 2010; see 2.1.2 for detail) was conducted. Two speakers of each accent took park in the Diapix

[6] Five Korean subjects reported that they had also lived in other English-speaking countries (e.g., the U.S, Australia, and Ireland) as adults for between 2 months and 3 years, but they were living in London at test and thus had been exposed to Standard Southern British English.

task together. It was conducted in two sound-treated booths that were specially set up for the Diapix task; each person was sitting in each booth facing each other through a large window, but they were not able to see each other's picture. To elicit casual speech, the task was performed in a normal listening condition (i.e., "no-barrier" condition). That is, speakers were able to hear each other clearly through microphone headsets (Beyerdynamic DT297). Their speech was recorded via the same headsets with 44100 16-bit samples per second. Each pair of speakers completed 8 scenes, each of which lasted an average of 11 minutes. The spontaneous speech obtained was then edited such that one stimulus comprised a section of speech produced by one talker describing a specific part of a DiapixUK scene, as shown in Table 2-1. Each of these stimuli was four to five seconds long on average.

Table 2-1: Examples of spontaneous speech stimuli from the Diapix recordings (SSBE: Standard Southern British English, KE: Korean-accented English, FE: Finnish-accented English).

DiapixUK picture[7]	Accent	Utterances
Farm 4 (Washing line)	SSBE	*We've got a washing line, with two white sheets on it.*
Farm 4 (Bowling)	KE	*There is one man who is just standing, holding the ball.*
Farm 4 (Beekeeper)	FE	*Mine is jumping in the air, and there's bees all around him.*
Beach 3 (Rocks)	SSBE	*One of the smaller rocks is a lighter brown.*
Beach 3 (Car)	KE	*I think the car was stuck, maybe stuck in the sand.*
Beach 3 (Rubbish bin)	FE	*Mine is like fallen over, and the lid is kind of open.*

For the read speech condition, the speakers were recorded reading the Basic English Lexicon (BEL) sentences (Calandruccio & Smiljanic, 2012), which are sentence materials developed for non-native speakers. Sentences were recorded for each

[7] These DiapixUK pictures are displayed in Figure 2-1. Specific scenes being described are shown in brackets.

speaker individually in a sound-treated booth via the same headset microphone (Beyerdynamic DT297) with 44100 16-bit samples per second.

In order to mix the stimuli with noise, speech-shaped noise was generated for each talker and speaking style; a smoothed long-term average spectrum was calculated using their recordings, and noise was then created such that it has the same spectral content as the spectrum. Prior to conducting the main experiment, 6 Korean listeners and 3 English listeners participated in a pilot study with stimuli mixed with various signal-to-noise ratios; quiet, 3 dB, -3 dB, -6 dB, -9 dB. Each subject participated in two of these SNR conditions and the quiet condition, which were presented in separate blocks. Stimuli for all three accents were mixed together within each block (see Chapter 2.2.3 for details of the procedure). This pilot study was conducted for exploratory purposes to determine appropriate levels of background noise for the main experiment. Statistical analyses were not performed for this pilot data because there were not enough subjects for each noise-level condition. The results were also averaged across different accents.

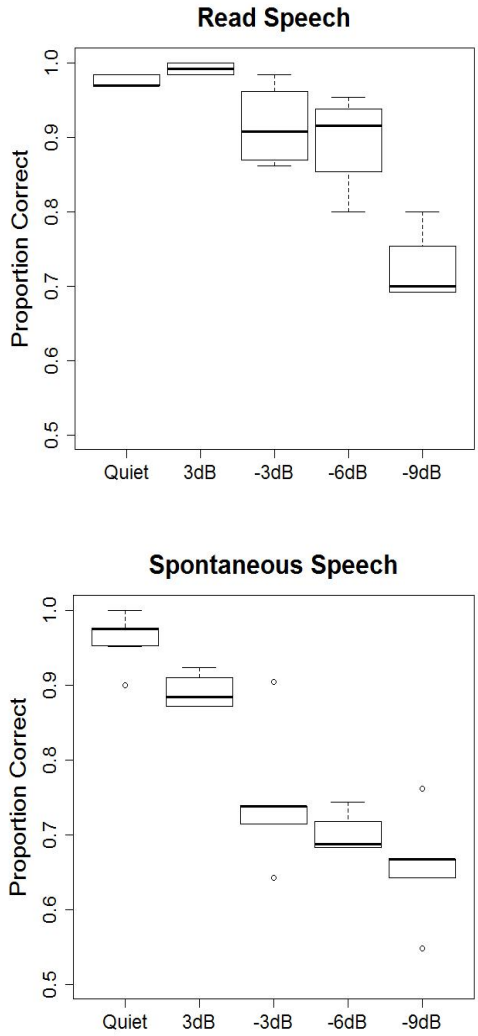
As displayed in Figure 2-2, the results indicated that the intelligibility of read speech materials was very high up until the noise level of -6 dB with the average proportion correct ranging between 0.88 and 0.99. The intelligibility level then decreased significantly at -9 dB with the mean accuracy of 0.72. This level of intelligibility was found at lower noise levels for spontaneous speech: -3 dB and -6 dB SNRs (i.e., mean accuracy was 0.75 and 0.70, respectively). This pattern of intelligibility was similarly found for native and non-native listeners (Table 2-2). Based on these results, different

SNRs were chosen for each speaking style condition for the main experiment, in order to achieve similar intelligibility levels: -8 dB for read speech and -4 dB for spontaneous speech. A single signal-to-noise ratio had to be used for each condition because using varying noise levels was infeasible given the difficulty of extracting appropriate stimuli from spontaneous speech.

Table 2-2: Pilot results – average speech recognition accuracy (i.e., proportion of correct responses) by signal-to-noise ratios for each speaking style and listener group. The results were averaged over all accents.

		Quiet	3dB	-3dB	-6dB	-9dB
Read speech	English listeners	0.97	0.99	0.98	0.92	0.80
	Korean listeners	0.98	0.99	0.89	0.88	0.70
	Mean	0.98	0.99	0.92	0.90	0.72
Spontaneous speech	English listeners	0.97	0.90	0.90	0.69	0.76
	Korean listeners	0.96	0.88	0.71	0.71	0.63
	Mean	0.96	0.89	0.75	0.70	0.66

Figure 2-2: Pilot results – speech recognition accuracy (i.e., proportion of correct responses) by signal-to-noise ratios for each speaking style, averaged over all listener groups and accents (top: read speech condition, bottom: spontaneous speech condition)



2.2.3 Procedure

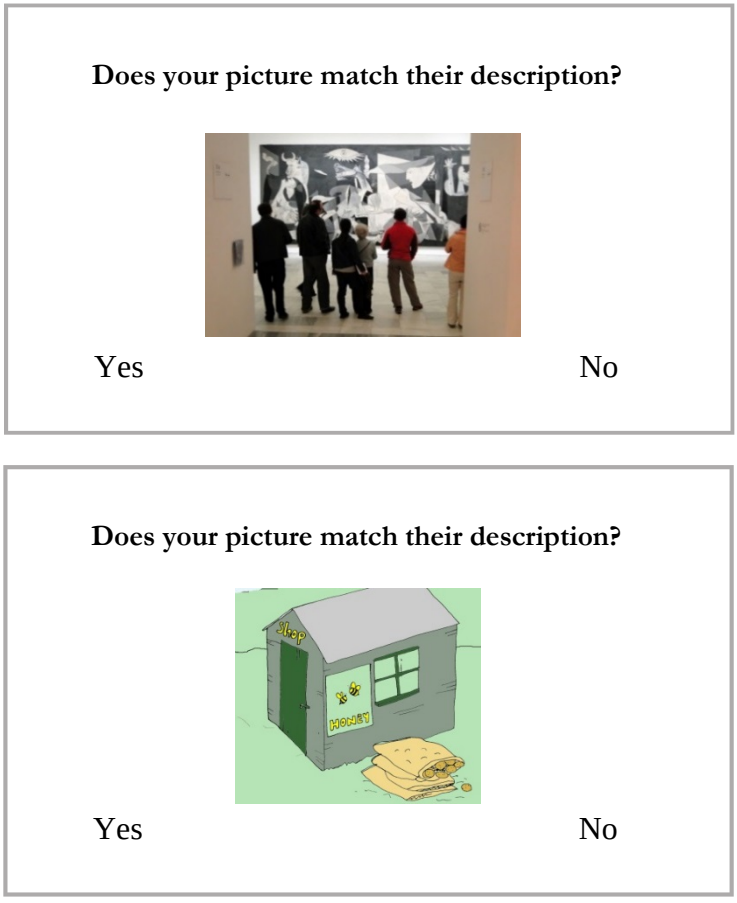
2.2.3.1 Speech recognition task

Subjects participated in a speech recognition task in which they listened to 195 read sentences which were mixed with the speech-shaped noise at the SNR of -8 dB, and 126 stimuli of spontaneous speech which were mixed with the noise at the SNR of -4 dB. The experiment consisted of a total of six blocks (i.e., 3 blocks for each speaking style condition). Stimuli for each of the three accent conditions were mixed together within each block to avoid any accent adaptation effects. The sentence assignment to different accents was counterbalanced between subjects, and each stimulus was presented only once. In the spontaneous speech condition, utterances describing each specific part of Diapix scenes (e.g., seesaw in a beach scene) were counterbalanced across subjects in a similar way. The utterances were not repeated. Within each block, half of the trials displayed pictures that matched what was being described in the speech stimuli and the other half displayed non-matching pictures. The order of stimuli was randomised within each block for each subject.

The experiment was performed via Praat (Boersma & Weenink, 2014). After hearing each stimulus, subjects had to decide whether what they heard matched the picture shown on the screen and click the right button (yes or no), as displayed in Figure 2-3. For the read speech condition, photos that either match the content of the target sentence or contain irrelevant content were displayed. For the spontaneous speech condition, parts of DiapixUK picture scenes were cropped and used for the experiment; pictures were chosen such that they either match or do not match the description given by the speaker. Pictures were carefully selected such that the difficulty of the task was

similar across stimuli. Specifically, the match between a speech stimulus and a picture was not based on something too trivial in the picture. Using the DiapixUK task also allowed for maintaining some control over linguistic content of speech such as lexical items. Additional care has been taken to select utterances that do not contain idiomatic expressions or complicated syntactic structures etc.

Figure 2-3: Schematic representation of the picture evaluation task (top: read speech condition, bottom: spontaneous speech condition). Listeners were asked to decide whether what they heard matched the picture shown on the screen and click yes or no.



2.2.3.2 ACCDIST

After finishing the picture evaluation task, Korean subjects were asked to take part in an additional recording session. They were recorded reading sentences from the BEL corpus (Calandruccio & Smiljanic, 2012) so that their accent could be analysed. Specifically, the acoustic similarity between the accents of Korean subjects (i.e., Korean listeners) and three groups of talkers (English, Finnish and Korean) was measured using a computation method called ACCDIST (Huckvale, 2004; 2007a,b). To this end, thirty BEL sentences read by Korean listeners and all six talkers were phonetically transcribed; automatic alignment was first performed using a forced aligner based on the HTK Hidden Markov Modelling Toolkit (1989), and it was then manually checked and corrected. The current study measured accent similarity based on vowel spectra and vowel duration which were previously shown to be effective in assessing accent-related differences across native and non-native accents (Pinet et al., 2011). The ACCDIST analysis was only conducted between Korean listeners and three groups of talkers in the present study, because English subjects (i.e., listeners) did not participate in this recording session.

To measure accent similarity based on vowel spectra, mel-frequency cepstral coefficients (MFCCs) vectors were calculated for the first and second half of each vowel segment. The Euclidean distance was then calculated between MFCC vectors of different instances of vowels (i.e., the same phonemes in different words were treated as distinct) within a single speaker to measure the vowel spectral contrasts that the individual speaker made. This process was used as a normalisation procedure to minimise the effects of individual speaker characteristics such as voice characteristics

that are not related to accent. These intra-speaker acoustic distances were then compared with those of another speaker for each pair of vowel instances to calculate the correlation between the two speakers. Accent similarity based on vowel duration was evaluated similarly, but by directly comparing between two speakers without computing intra-speaker acoustic distances (Huckvale, 2004; 2007a,b).

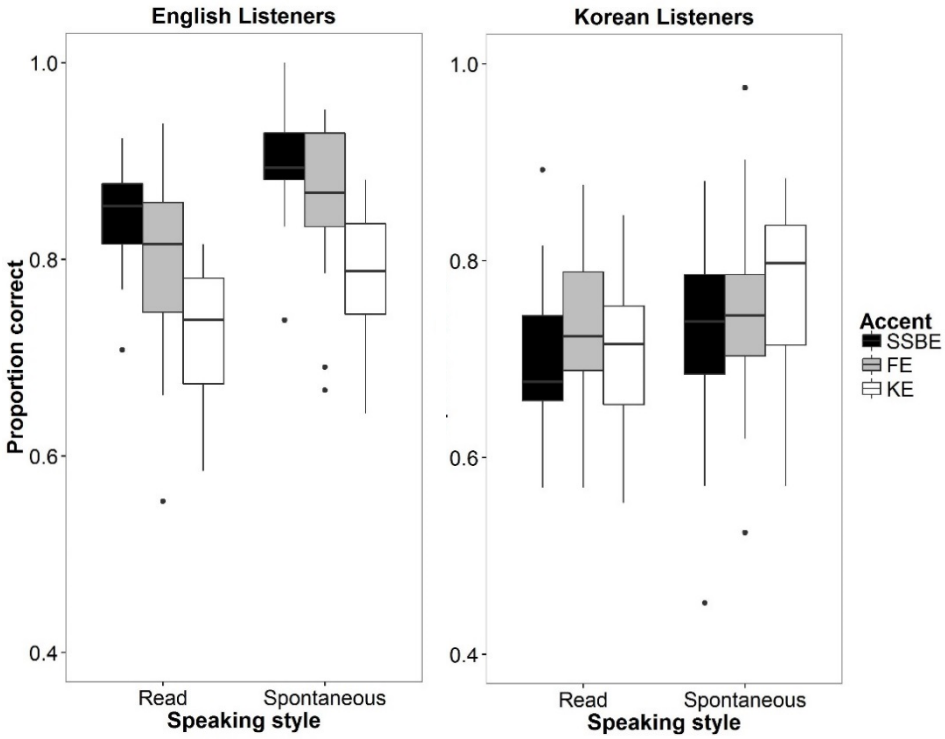
2.3 Results

A logistic mixed-effects analysis was conducted in R using the package lme4 (Bates, Mächler, Bolker, & Walker, 2015), with talker accent (English, Finnish, Korean), listener group (English, Korean) and speaking style (read speech, spontaneous speech) included as fixed effects, and with by-subject and by-stimuli random intercepts. Because this study was interested in examining all the main effects and interactions of these fixed factors, they were all included in the model rather than building a model that best fit the data. The dependent variable was the response in the speech-in-noise recognition task (i.e., correct or incorrect). The package CAR (Fox & Weisberg, 2002) was used to calculate type II analysis-of-variance tables. The multcomp package (Hothorn, Bretz, & Westfall, 2008) was used to run post-hoc analyses.

Figure 2-4 displays the mean proportion correct for each listener group and accent condition. Overall, it appears that Korean listeners had lower recognition accuracy than English listeners; there was a significant main effect of listener group, $\chi^2(1) = 26.55$, $p < 0.001$. However, the two-way interaction between talker accent and listener group was also significant, $\chi^2(2) = 90.28$, $p < 0.001$, suggesting a clear talker-listener accent interaction. Specifically, English listeners showed the highest recognition

performance on the native accent (i.e., SSBE), followed by Finnish-accented English (i.e., FE), and the lowest performance on Korean-accented English (i.e., KE). Tukey post-hoc tests using the multcomp package confirmed that the differences between all three accents were significant for English listeners, $p < 0.001$. In contrast, the recognition accuracy of Korean listeners was not significantly different for the different accents overall (Tukey post-hoc results: $p = 0.273$ for SSBE and FE; $p = 0.997$ for FE and KE; $p = 0.540$ for SSBE and KE). The main effect of talker accent was also significant, $\chi^2(2)= 25.45$, $p < 0.001$[8].

Figure 2-4: Speech-in-noise recognition accuracy of English and Korean listeners by speaking style and speaker accent (SSBE: Standard Southern British English, FE: Finnish-accented English, KE: Korean-accented English)



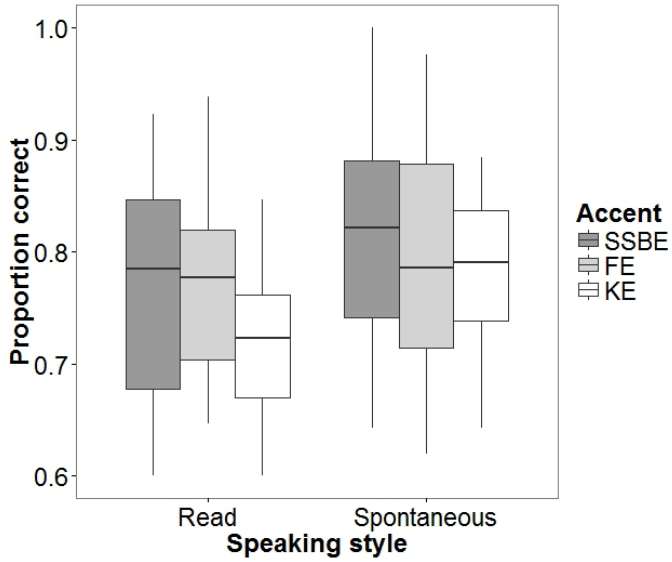
[8] Tukey post-hoc tests found that Korean-accented speech was significantly less intelligible than Standard Southern British English, $p < 0.001$, and the difference between Korean-accented speech and Finnish-accented speech was marginally significant, $p = 0.0676$. The difference between Standard Southern British English and Finnish-accented speech was not significant, $p = 0.2694$.

In addition, the main effect of speaking style was significant, $\chi^2(1)= 12.61$, $p <0.001$. That is, recognition accuracy was higher for spontaneous speech (i.e., casual speech) than for read speech overall because of the higher signal-to-noise ratio used for the spontaneous speech condition. This indicates that the two different SNR levels did not completely equalise performance levels in the two conditions. The two-way interaction between speaking style and listener group was significant, $\chi^2(1)= 9.30$, $p = 0.0023$. Specifically, the difference in performance between English and Korean listeners was larger in the spontaneous speech condition ($M_{English} = 0.85$; $M_{Korean} = 0.75$) than in the read speech condition ($M_{English} = 0.78$; $M_{Korean} = 0.71$). Despite the fact that the lower noise level used in the spontaneous speech condition increased overall intelligibility compared to the read speech condition, the improvement was smaller for Korean than English listeners.

Furthermore, as shown in Figure 2-4, the speech recognition performance of English listeners was affected by accents similarly in read and spontaneous speech conditions, with SSBE being most intelligible, followed by FE, and KE being least intelligible. However, Korean listeners showed some indication of a trend for them to understand Korean-accented speech better than other accents in the spontaneous speech condition, although the three-way interaction between speaking style, talker accent and listener group did not reach significance, $p = 0.0978$. However, there was a significant interaction between talker accent and speaking style, $\chi^2(2)= 6.55$, $p = 0.0379$. Tukey post-hoc tests showed that accent differences were only found in the read speech condition. Specifically, both SSBE and FE were significantly more intelligible than KE in the read speech condition, $p < 0.001$, but none of the comparisons were

significant in the spontaneous speech condition, $p > 0.05$ (Figure 2-5). It should also be noted that the best-fitting model without the insignificant three-way interaction (speaking style * talker accent * listener group) also produced the same significant effects.

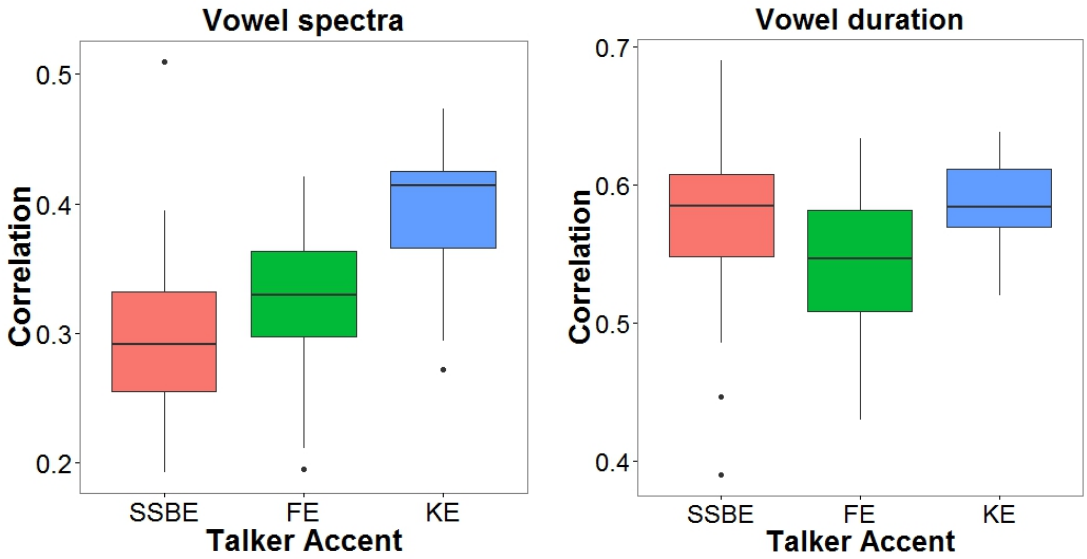
Figure 2-5: Speech-in-noise recognition accuracy by speaking style and speaker accent averaged over all listeners (SSBE: Standard Southern British English, FE: Finnish-accented English, KE: Korean-accented English)



As explained previously, the accent of each Korean subject (i.e., listener) was compared to each of the six talkers using the ACCDIST metric. Separate linear mixed-effects analyses were performed for vowel spectra and vowel duration measurements; in each model, ACCDIST (i.e., correlation coefficients) for each talker-listener pair was included as a dependent variable, talker accent (SSBE, FE, KE) as fixed effects, and with a by-subject (i.e., listeners) random intercept. For vowel spectra measurements, there was a main effect of talker accent, $\chi^2(2) = 50.95$, $p < 0.001$. As displayed in Figure 2-6, Tukey post-hoc analyses found that the Korean listeners'

accents were closer in vowel spectral qualities to Korean-accented English than to Standard Southern British or Finnish-accented English, $p < 0.001$, with no significant difference between the distance to SSBE and FE, $p = 0.237$. The main effect of talker accent was also significant for vowel duration, $\chi^2(2) = 13.26$, $p = 0.0013$. As shown in Figure 2-6, Tukey post-hoc tests revealed that the Korean listeners' accents were equally similar to KE and SSBE, $p = 0.5382$, but were significantly more distant from FE ($p = 0.0345$ and $p = 0.0011$, when compared to SSBE and KE, respectively).

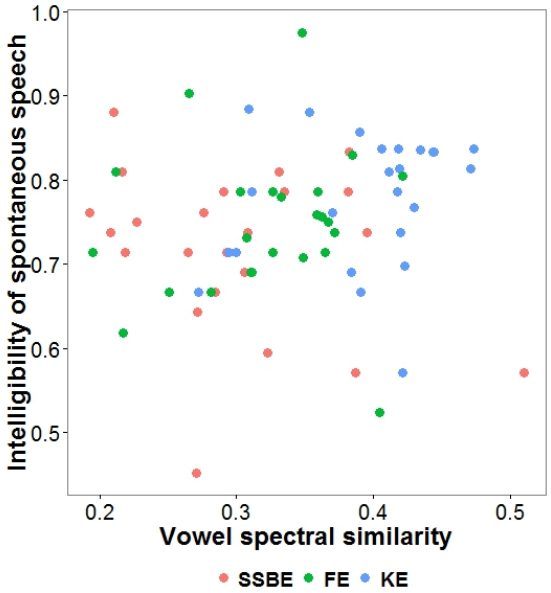
Figure 2-6: Accent similarity (i.e., accent correlation) between Korean listeners and three groups of talkers (SSBE: Standard Southern British English, FE: Finnish-accented English, KE: Korean-accented English) in terms of vowel spectral qualities and duration.



To investigate links between accent similarity and speech recognition performance, additional linear mixed-effects models were performed with average proportion correct used as the dependent variable, the level of accent similarity based on each of the two ACCDIST measures (i.e., averaged by talker accent for each listener) as fixed effects and with a by-subject random intercept. Because the ACCDIST analysis was

conducted using 30 BEL sentences, there were no accent similarity data for spontaneous casual speech. Separate linear mixed-effects analyses were thus carried out for the speech recognition accuracy of read and casual speech conditions. The results showed that accent similarity measures based on vowel spectra and vowel duration were both not significant predictors of the intelligibility of read speech, $p = 0.7872$ and $p = 0.5705$, respectively[9]. When the proportion correct for spontaneous speech was used as the dependent variable, the main effect of vowel spectral quality was marginally significant, $\chi^2(1) = 3.21$, $p = 0.0733$ (Figure 2-7), and the main effect of vowel duration was not significant, $p = 0.3459$. The direct relationship between accent similarity and intelligibility thus appears to be less strong for Korean listeners of the current study than previously found (Pinet et al., 2011).

Figure 2-7: Scatterplot of Korean listeners' accent similarity to each of the talker accents based on vowel spectra (x-axis) vs their recognition acccuracy (i.e., proportion correct; y-axis) for these accents in the spontaneous speech condition (SSBE: Standard Southern British English, FE: Finnish-accented English, KE: Korean-accented English).



[9] When the proportion correct was averaged across different speaking styles, accent distances based on both vowel spectra and duration were not significant.

The similarity between the accents of the six talkers was also examined for exploratory purposes; because there were only two speakers for each accent (i.e., four pairs of speakers for each accent comparison), statistical tests were not carried out. Table 2-3 displays the results of the ACCDIST analysis averaged for each talker accent pair. Overall, it seems that the accents of Southern British and Finnish speakers were closest to each other both in terms of vowel spectral qualities and vowel duration, whereas the accent of Korean speakers was most distant from the two accents.

Table 2-3: Descriptive statistics for the ACCDIST analysis conducted across talker accents

Accent similarity (i.e., accent correlations) across talker accents	Vowel spectra Mean(sd)	Vowel duration Mean(sd)
Standard Southern British - Finnish	0.4577 (0.0705)	0.6781 (0.0189)
Standard Southern British - Korean	0.3351 (0.0341)	0.5073 (0.0584)
Finnish - Korean	0.3180 (0.0099)	0.5316 (0.0357)

2.4 Discussion

The current study investigated how speech recognition by native and non-native listeners is modulated by speech style (spontaneous casual vs. read) and the accents of the talkers and listeners in the presence of background noise. A new speech recognition task (i.e., picture evaluation task) was developed to evaluate speech recognition performance for both read and spontaneous speech. The results demonstrated that native listeners were more accurate at the speech-in-noise recognition task overall than were non-native listeners. In addition, non-native listeners suffered more from casual speech than did native listeners. There was also a clear interaction between the accents of the talker and the listener across speaking styles, supporting previous work (e.g.,

Bent & Bradlow, 2003). Native English listeners showed a clear intelligibility benefit for their own accent; they found Standard Southern British English most intelligible, followed by Finnish-accented English; Korean-accented English was the least intelligible to them. In contrast, Korean listeners found all three accents similarly intelligible, suggesting a relative intelligibility advantage for Korean-accented English, which was the least intelligible accent for English listeners.

As expected, the performance of native listeners was more accurate than that of non-native listeners overall, but the recognition accuracy of the two listener groups was differentially modulated by speech style and accent. First, there was a significant interaction between speech style and listener group. Specifically, the difference in speech recognition accuracy between English and Korean listeners was larger in the spontaneous speech condition than in the read speech condition. In other words, due to the less-severe noise level used in the spontaneous speech condition, the intelligibility of spontaneous speech was higher than that of read speech, but this intelligibility gain was greater for native than non-native listeners. This occurred likely because features of spontaneous casual speech such as reduction phenomena were more deleterious to non-native listeners, thereby attenuating the benefit from listening in a less-severe noise condition.

One may argue that this result occurred because a higher degree of signal clarity is required than that provided in the -4 dB SNR condition for non-native listeners' speech recognition to improve (i.e., intelligibility could increase at different rates for native and non-native listeners with decreasing levels of background noise). That is, the

effects of speaking style seen in this study could have been partially driven by different levels of noise. However, this seems less likely given the results of the pilot study; Korean listeners' recognition performance for read speech improved as much as that of English listeners when the noise level decreased from -9 dB to -3 dB (the intelligibility gain was 19 and 18 percentage points, respectively).

Furthermore, irrespective of speaking style, English listeners displayed a clear advantage for their own accent (i.e., SSBE). They also found Finnish-accented English more intelligible than Korean-accented English. The ACCDIST analysis performed across pairs of talkers can account for the intelligibility differences between these accents; the accent distance between Finnish and English talkers was closest, and the accent of Korean talkers was more distant from the two accents. That is, the Finnish talkers had an accent that was more similar to Standard Southern British English, suggesting that they were more proficient L2 speakers of English who had acquired more native-like phonetic/phonological representations of English compared to the Korean talkers. It thus appears that it was easier for English listeners (i.e., SSBE speakers) to map the Finnish-accented speech onto their phonetic and phonological representations, thereby finding the accent easier to understand than the Korean accent that was acoustically more distant.

The recognition accuracy of Korean listeners was modulated by speaker accent in a more complex way. The accent similarity results revealed that the accent of Korean listeners resembled the accent of Korean talkers more than that of English or Finnish talkers in terms of vowel spectra. The accent similarity results based on vowel duration

showed that Korean listeners were equally similar to Korean and English talkers and less similar to Finnish talkers. Overall, it is apparent that the accent of Korean listeners was acoustically closest to the accent of Korean talkers, indicating that they had more similar phonetic and phonological representations of the L2 (i.e., their L2 knowledge was influenced by the same L1), and were at similar stages of L2 acquisition. Nonetheless, the Korean listeners found the other accents as intelligible as the Korean accent. Likewise, accent similarity was not a significant predictor of intelligibility for Korean listeners in the mixed-model analysis.

It thus seems that there were other factors determining accent intelligibility for Korean listeners. It is possible that accent familiarity played a role in modulating accent intelligibility (e.g., Adank et al., 2009). That is, SSBE and KE were equally intelligible to Korean listeners possibly because they had been living in England at the time of testing and were thus familiar with the Southern British accent to some extent. Specifically, Korean listeners may have learned to map the native accent onto their underlying phonological representations through experience with that accent, despite not having fully developed native-like underlying representations (e.g., Sumner & Samuel, 2009). However, Finnish-accented English was also equally intelligible as the other accents despite the fact that the Korean listeners had little exposure to Finnish-accented English. Although this might seem to invalidate the familiarity account, it is possible that Korean listeners were recruiting representations that they had developed through exposure to SSBE to understand the Finnish speakers, whose accents were acoustically fairly similar to SSBE. For example, according to exemplar theories of speech perception, listeners process speech sounds by matching them with acoustically

similar exemplars that they have previously encountered (e.g., Goldinger, 1998; Goldinger, 1996; Johnson, 1997).

It should also be noted that non-native listeners' perceptual and phonological processes that are used for speech recognition might not necessarily match their production accuracy. The Speech Learning Model (Flege, 1995) predicts that the production of L2 sounds depends on how accurately the sounds are perceived, and the production and perception of L2 sounds have been shown to be moderately correlated in several studies (e.g., Flege, 1993; Flege, Bohn, & Jang, 1997; Flege, MacKay, & Meador, 1999). However, they do not necessarily develop in parallel; acquiring production skills may require more extensive early language experience (e.g., experience of speaking as well as hearing) than acquiring perception skills (Oh, Jun, Knightly, & Au, 2003), and conversely, production can also precede perception in some cases (e.g., Sheldon & Strange, 1982 for Japanese speakers learning English /r/ and /l/; Evans & Iverson, 2007 for native listeners adapting to a new accent). Although it is hard to further interpret this result without comparing the perception and production for specific L2 sounds, it is plausible that Korean listeners have developed their underlying phonological processes to understand native or experienced L2 speakers to a certain level, but their production has not reached the same level of proficiency.

Alternatively, it is possible that the English and Finnish talkers inherently had higher intelligibility compared to the Korean talkers regardless of their accent-related features, which may have helped the Korean listeners overcome the relatively difficulty in understanding the accents that were different from their own. Previous

research has shown that intelligibility is also determined by inherent talker clarity that is related to global acoustic-phonetic characteristics such as energy in the mid-frequency region, and the intelligibility of individual talkers has been shown to be highly correlated across native and non-native listeners (van Dommelen & Hazan, 2012; Iverson, Pinet, & Evans, 2014). The effect of individual talker intelligibility is difficult to know with the current number of talkers, but it would be interesting to explore its effect in future studies.

Although there was an indication that Korean listeners had higher intelligibility for their own accent in the casual speech condition (Figure 2-4), this was only marginally significant. Instead, there was a significant two-way interaction between talker accent and speaking style. Specifically, intelligibility differences between accents, with SSBE and FE being more intelligible than KE, were only found in the read speech condition (Figure 2-5). It is possible that the casual speech of the less-fluent Korean talkers was relatively more intelligible because it had fewer reduced pronunciation variants (e.g., fewer casual speech processes or more salient acoustic-phonetic cues such as word-final stop releases, Bent & Bradlow, 2003) compared to that of native or Finnish talkers, thereby diminishing the intelligibility differences between the accents in the spontaneous speech condition. Non-native speakers who are not at an advanced stage of the target language acquisition are more likely to produce canonical forms even though they are deviant from the native norm, and this can help the listener to overcome the difficulty in understanding the L2 accent that is distant from their own.

In summary, the results of the current study support the previous finding that speech recognition in noise is affected by the accents of the talkers and listeners, with English listeners showing a distinct advantage for their own accent, and Korean listeners finding all three accents similarly intelligible. Accent similarity was able to account for the differences in accent intelligibility for English listeners (e.g., Finnish speakers were more intelligible than Korean speakers), but it appears that other factors may have also contributed to the observed patterns of accent intelligibility for Korean listeners. Their experience with SSBE might have helped them to understand the native speakers and the Finnish speakers who had more native-like accents, even though their production skills did not match those of the speakers. It may be advantageous to test non-native listeners and talkers with varying levels of L2 proficiency to clarify these findings because it is possible that the proficiency of the Korean listeners somewhat varied in this study, causing mixed accent effects.

In addition, the new speech recognition task (i.e., picture evaluation task) was found to be successful in evaluating speech recognition performance for both spontaneous and read speech materials and thus has great potential for use in future research to examine the recognition of spontaneous speech. The current findings also demonstrate that talker-listener accent interactions that had previously been found using read speech extend to spontaneous casual speech, and that features of casual speech can make L2 speech perception more challenging. However, it should be noted that the effect of speaking style was not completely isolated from that of noise in the current study as different SNR levels were used for each speaking style condition. It would

thus be ideal to have the same noise levels between the two conditions in a future study.

Lastly, the current findings did not clearly support the prediction that talker-listener accent interactions could be stronger in natural communicative situations where speakers talk casually. That being said, the intelligibility differences between the accents found in the read speech condition (i.e., Korean talkers were less intelligible than English and Finnish talkers) disappeared in the spontaneous speech condition, suggesting that certain characteristics of casual speech produced by inexperienced non-native talkers could in fact be beneficial for the listener regardless of what accent the listener has (e.g., less reduced forms). It remains for future research to investigate which processes or acoustic-phonetic characteristics in non-native casual speech lead to increased intelligibility, and how they affect speech recognition by native and non-native listeners. It would be interesting to further investigate these questions using spontaneous speech to understand how everyday speech communication is influenced by speaker accent.

Chapter 3 Cortical entrainment to the amplitude envelope of speech

3.1 Introduction

Much of our knowledge of L2 speech perception has been based on findings from experiments that used isolated words or syllables as stimuli. However, speech comprehension in everyday life involves processing speech that is longer (i.e., continuous speech) as well as more phonetically variable. The current study originally set out to develop methods that measure how L2 listeners process continuous speech. The processing of longer utterances involves a whole range of speech recognition processes such as phoneme recognition, word segmentation, and lexical access, but recent research has shown that speech recognition also involves processing slow amplitude fluctuations in speech (i.e., the amplitude envelope) at the cortical level (e.g., Ahissar et al., 2001; Luo & Poeppel, 2007; Peelle et al., 2013) which has not been investigated in L2 speech perception research. This cortical response to speech may help account for speech recognition difficulties experienced by L2 listeners in everyday life that are related to processing connected speech.

The present study thus examined neural entrainment to the amplitude envelope of speech using EEG, while subjects listened to continuous stories in their native language, second language or a language that they did not understand. This measure was chosen because previous research has suggested that there is a positive relationship between neural tracking of the temporal envelope and speech comprehension (e.g., Peelle et al., 2013). That is, one could expect that native listeners have greater entrainment to the speech envelope that do non-native listeners; this could

provide evidence supporting the hypothesis that the effect of language experience is seen at an early, auditory level of speech processing, and that this effect extends to the processing of continuous speech. Using a cross-linguistic design, the present study was also able to investigate the much-debated issue on the link between entrainment and speech intelligibility (e.g., Peelle et al., 2013; Howard & Poeppel, 2010; Millman, Johnson, & Prendergast, 2015) without altering the acoustic properties of the speech signals.

3.1.1 Cortical entrainment to the amplitude envelope of speech

An increasing number of studies have shown that low frequency neural oscillations in the auditory cortex (1~8 Hz) become phase-locked to slow temporal fluctuations in the speech signal, that is, the amplitude envelope, during speech perception (e.g., Ahissar et al., 2001; Luo & Poeppel, 2007; Peelle et al., 2013). This neural activity is also referred to as 'cortical/neural entrainment to speech'. Temporal modulations particularly in the theta band (4-8 Hz) are related to the production of syllables (i.e., the mean syllable duration of English is approximately 200 milliseconds in spontaneous speech; Greenberg, 1999). Entraining to those quasi-rhythmic amplitude fluctuations in speech is thought of as increasing the efficiency of speech processing; the phase of neural oscillations is modulated such that critical acoustic information delivered by slow amplitude fluctuations arrives at a time of high neural excitability (possibly via phase-resetting by stimulus onsets; Peelle & Davis, 2012).

The amplitude envelope in the speech signal is used to carry important linguistic information - both segmental (e.g., voicing and manner of articulation) and

suprasegmental cues (e.g., stress; Rosen, 1992). Evidence from behavioural studies suggests that listeners are thus able to understand speech fairly well with limited spectral information if low-frequency envelope cues are available (e.g., Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Similarly, Ghitza and Greenberg (2009) found that while intelligibility was poor when listening to time-compressed speech (compressed by a factor of 3), intelligibility increased dramatically when fragments of silence were inserted, such that the duration of a compressed portion of speech and added silence matched the duration of the original syllable (i.e., when the original rate of low-frequency information was preserved). These studies demonstrate that the low-frequency envelope information is important for speech recognition.

Recent neuroimaging evidence has suggested that low-frequency entrainment to speech does not merely reflect neural encoding of the speech envelope, but is associated with other important processes of speech recognition. Several studies have found that the degree of phase-locking measured between neural signals and amplitude envelopes is correlated with speech intelligibility (e.g., Peelle et al., 2013; see Chapter 3.1.2 for details). In addition, in a competing-talker background, attention can modulate envelope tracking activity, selectively enhancing the entrainment to the target talker (e.g., Ding & Simon, 2012a; Kerlin, Shahin, & Miller, 2010). Speech entrainment is thus thought to reflect speech segregation, a prerequisite for successful speech recognition in complex auditory scenes (e.g., Ding & Simon, 2012a; see Chapter 4).

The mechanism underlying neural entrainment to speech has been elaborated by Giraud and Poeppel (2012) who hypothesized that delta (1-3 Hz), theta (4-8 Hz), and gamma (low gamma; 25-35 Hz) oscillations track multi-timescale units of speech; namely, lexical and phrasal units (that form a prosodic unit such as 'Intonational Phrase' carrying an intonation contour), syllables, and phonemes, respectively. This theory is largely based on the close temporal correspondence between these speech units and brain rhythms. As previously mentioned, syllables are produced at similar rates to theta oscillations; lexical and phrasal units occur at slower rates around 1-2 Hz (delta); and cues used to differentiate phonemes (e.g., segmental cues to manner of articulation or voicing) occur at faster modulation rates roughly matching the gamma band. The authors thus theorised that this speech-brain alignment could be a neural mechanism of speech processing which allows for segmentation of connected speech into discrete units[10]. Gross et al. (2013) also showed this multi-timescale neural tracking of speech in their magnetoencephalography (MEG) study; the phase of delta and theta oscillations and the amplitude of gamma oscillations (35-45 Hz) were entrained to the amplitude envelope of speech. In addition, cross-frequency coupling between neural oscillations in theta and gamma bands (i.e., theta-gamma nesting) is thought to be one of the principles in this oscillation-based speech processing (Giraud & Poeppel, 2012; Ghitza, 2011); the phase of theta determines the properties (i.e., amplitude) of gamma oscillations[11].

[10] Ghitza (2011) also proposed a parallel model 'Tempo' which postulates that a cascaded array of oscillators tracks the rhythm of speech, which governs the decoding time during memory access.

[11] This theta-gamma nesting has been demonstrated in an earlier study by Lakatos et al. (2005) which found that neural oscillations occurring in different frequency ranges are hierarchically organised; delta phase modulates the amplitude of theta oscillations, and theta phase modulates the amplitude of gamma oscillations ('oscillatory hierarchy hypothesis').

Although speech entrainment has been found at multiple time scales (e.g., Gross et al., 2013), it remains to be seen whether it reflects speech processing at linguistic levels (i.e., phonological, lexical, semantic and syntactic). This somewhat overlooked question seems to be important because these speech units (i.e., phonemes, syllables, words, and phrases) are segmented by using language-specific linguistic knowledge rather than purely relying on acoustic cues (see Ding, Melloni, Zhang, Tian, & Poeppel, 2016). Similarly, "syllable" is a phonological unit that is defined differently depending on language (i.e., languages have different syllable structures), and the amplitude envelope does not necessarily provide clear information about syllables (Cummins, 2012). In this sense, theta-band entrainment to the amplitude envelope does not necessarily mean that listeners are parsing "syllables". It thus remains to be seen how much speech entrainment is associated with "linguistic" processing (see Obleser, Herrmann, & Henry, 2012 for other problems of the theory).

3.1.2 Entrainment to the speech envelope and speech intelligibility

While a growing body of literature has demonstrated that neural oscillations track slow amplitude modulations in the speech signal during speech perception, it is controversial whether this neural activity purely reflects auditory processing of speech or is related to higher-level linguistic processing. More specifically, the relationship between speech intelligibility and cortical entrainment to the temporal envelope has been extensively studied in recent papers which have produced inconsistent findings. This question is closely related to the aim of the present study which was to examine the role of native language experience on auditory cortical processing of the temporal envelope. This section will discuss the results from the previous studies in detail. It is

important to note that all these studies used altered speech signals such as vocoded speech (Peelle et al., 2013; Ding, Chatterjee, & Simon, 2014), time-compressed speech (Ahissar et al., 2001; Nourski et al., 2009) and time-reversed speech (e.g., Howard & Poeppel, 2010; Gross et al., 2013) or added background noise (Ding & Simon, 2013 to manipulate speech intelligibility.

Table 3-1: Studies that examined the relationship between speech intelligibility and neural entrainment to the temporal envelope

Study	Stimuli	Positive relationship found?
Ahissar et al. (2001)	Time-compressed speech	Yes
Luo & Poeppel (2007)	Speech-noise chimaeras	Yes
Peelle et al. (2013)	Noise-vocoded speech	Yes
Gross et al. (2013)	Backward-presented speech	Yes
Ding & Simon (2013)	Spectrally-matched background noise	Yes
Ding et al. (2014)	Noise-vocoded speech	Yes
Doelling, Arnal, Ghitza, & Poeppel (2014)	Envelope alterations[12]	Yes
Nourski at al. (2009)[13]	Time-compressed speech	No
Howard & Poeppel (2010)	Time-reversed speech	No
Millman et al. (2015)	Tone-vocoded speech before/after the presentation of the original sentence ("pop-out")	No

Some of the studies shown in Table 3-1 found greater entrainment to the speech envelope when the intelligibility of speech was greater. One of the earliest studies by Ahissar et al. (2001) examined cortical responses to time-compressed speech that was created with varying compression ratios (0.2, 0.35, 0.5 and 0.75) using MEG. The

[12] In this paper, the amplitude envelope was directly altered. When the amplitude envelope was distorted, entrainment to the envelope and intelligibility decreased.
[13] Nourski et al. (2009) reported mixed results as to the relationship between envelope tracking and speech intelligibility (see p.77 for details).

temporal envelope was distorted accordingly. They found that the degree of phase-locking and frequency-matching (computed from Fast Fourier Transforms of the signals) between the temporal envelopes of speech stimuli and neural responses was correlated with speech comprehension. They also found a significant individual-level correlation between speech comprehension and the degree of frequency-matching. They thus argued that neural phase-locking to the temporal envelope is a prerequisite for speech comprehension. Luo and Poeppel (2007) further investigated this issue, also in an MEG study, where the intelligibility of speech was manipulated using speech-noise chimaeras (i.e., 4-band chimaeras containing only envelope information and 1-band chimaeras containing only fine structure information). By calculating phase coherence between trials (i.e., neural signals) within and across sentences, they found that phase patterns in theta-band neural responses can discriminate between different sentences (i.e., greater coherence for within-stimulus trials), and that the accuracy of discrimination was correlated with speech intelligibility. In contrast, theta-band power was not able to discriminate between different sentences, suggesting that phase modulation is the key mechanism of theta-band envelope tracking.

Peelle et al. (2013) compared entrainment to noise-vocoded speech with different numbers of channels using MEG. The noise-vocoding technique was used to vary the amount of spectral detail in the speech signal while preserving the amplitude envelope, which is different from some of the previous studies, in which the amplitude envelope was directly altered as a result of the acoustic manipulation used (e.g., time-compression; Ahissar et al. 2001). They similarly found that neural phase-locking to speech was significantly greater for intelligible stimuli (i.e., 16 channel vocoded

sentences) than for unintelligible stimuli (i.e., 1 channel vocoded sentences). They also showed that this intelligibility effect was seen in the left hemisphere. Moreover, this occurred despite the fact that the amplitude envelope was more or less preserved in both conditions. The authors thus argued that it is likely because listeners were able to predict the onset of upcoming speech events (e.g., words) using linguistic information available in the speech signal. However, the difference between 4 channel vocoded (i.e., moderately intelligible) and 4 channel spectrally rotated (i.e., unintelligible) speech was less clear; they found increased entrainment for the intelligible condition only in the region of interest (ROI) analysis (i.e., for a 5 mm radius sphere centred on the middle temporal gyrus peak), not in the whole-brain analysis.

Gross et al. (2013) also found greater theta and delta entrainment to the envelope of normal speech (i.e., intelligible) than to that of the backward-played counterpart (i.e., unintelligible). Reversing speech in time can make it unintelligible (Saberi & Perrott, 1999), while preserving the overall properties of the amplitude envelope. Furthermore, their results showed that the amplitude of gamma oscillations was phase-locked to the speech envelope, and the degree of phase-locking was also greater for normal speech than for backward-presented speech. In addition, theta-gamma and delta-theta cross-frequency coupling (between theta phase and gamma amplitude, and delta phase and theta amplitude, respectively) was stronger for intelligible than unintelligible speech. It should be noted, however, that time-reversal can alter the amplitude envelope in some aspects (e.g., sounds containing acoustic transients such as plosives) and possibly affect the neural processing of the envelope (e.g., Peña & Melloni, 2012).

When background noise is added to speech, the results appear to be more complicated; Ding and Simon (2013) found that entrainment to slower temporal modulations in speech (< 4 Hz) was robust to background noise up to the signal to noise ratio (SNR) of -6 dB, whereas entrainment to faster modulations (4-8 Hz) decreased with decreasing SNR levels. There was also a significant positive correlation between subjectively rated intelligibility scores and entrainment accuracy for individual listeners at the SNR of -3 dB. Based on these results, the authors suggested that more precise encoding of the temporal envelope enhances speech comprehension in noisy environments, not the other way around as suggested in Peelle et al. (2013), and that the bottleneck for speech recognition in noise lies in the ability to extract the speech signal from background noise (i.e., auditory processing).

Moreover, Ding et al. (2014) directly investigated the effect of spectro-temporal fine structure on cortical entrainment to the speech envelope; half of their speech stimuli were mixed with spectrally matched stationary noise at the SNR of 3 dB, and the other half were not. Each of these stimuli (i.e., speech and speech-in-noise mixture) was then either noise-vocoded (4-channel and 8-channel) or unprocessed. When listeners heard natural, unprocessed speech, entrainment was robust to background noise, similar to what Ding and Simon (2013) found. However, as the spectral resolution of speech was reduced, entrainment to speech decreased in background noise in both delta (1~4 Hz) and theta (4~8 Hz) ranges. There were also significant positive correlations between individual listeners' speech intelligibility scores and delta-band entrainment in some of the conditions (i.e., 4-channel vocoded speech in quiet and 8-channel vocoded

speech in quiet and noise). That is, entrainment to the temporal envelope was modulated by the spectro-temporal fine structure of the speech signal.

The authors thus concluded that neural entrainment to the temporal envelope in fact reflects a collective, object-level neural representation of speech achieved by 'an analysis-by-synthesis process' (e.g., Poeppel, Idsardi, & van Wassenhove, 2008; Shamma, Elhilali, & Micheyl, 2011). In an analysis-by-synthesis process, multiple acoustic features of a complex auditory scene are first encoded sub-cortically in the analysis phase, and features belonging to a single auditory object are then grouped together using segregation cues in the synthesis phase. The authors suggested that entrainment to speech is thus enhanced when greater spectral cues are available to segregate the target auditory object from background noise. It should be noted, however, that the effect of noise vocoding was different in quiet; delta-band neural entrainment increased as the spectral resolution of speech decreased, whereas theta-band entrainment decreased. The authors argued that delta-band entrainment might reflect increased listening effort, because delta activity can reflect top-down attention (Schroeder & Lakatos, 2009).

Together, above-mentioned studies are all in agreement with one another in that they found that neural synchronisation to the amplitude envelope of speech was correlated with the amount of spectral detail in the speech signal or speech intelligibility. However, it remains unclear whether speech comprehension directly enhances tracking/encoding of low-frequency amplitude modulations (Peelle et al., 2013), or whether acoustic properties of the speech signal such as the amount of spectral detail

(e.g., Ding et al., 2014) or alterations in the amplitude envelope itself (e.g., time-reserved speech) affect neural phase-locking to the speech envelope, which in turn, affects later linguistic processing. If the latter is the case, speech comprehension is not directly related with envelope tracking.

Interestingly, some studies failed to find a positive relationship between entrainment and intelligibility as shown in Table 3-1. In Howard and Poeppel (2010), there was no significant effect of comprehension on phase-locked neural responses; the accuracy of sentence discrimination using theta-band phase patterns of single-trial MEG signals was not different for unintelligible, time-reversed sentences, or unprocessed counterparts, indicating that envelope tracking could be independent of speech comprehension. The authors concluded that theta-band speech tracking reflects cortical processing of low-frequency temporal modulations that are essential to intelligibility, but it does not directly reflect higher-level linguistic processing.

Nourski et al. (2009) reported somewhat mixed results regarding the relationship between speech entrainment and comprehension. They directly observed envelope tracking from Heschl's gyrus (HG) by measuring average evoked potentials (AEPs) and high-frequency power (70~250 Hz) in the electrocorticogram (ECoG)[14] in response to time-compressed speech. They found that the envelope-tracking response in the high-frequency activity of the ECoG was apparent in both left and right hemispheres even when the speech was made unintelligible at high compression rates (0.3 to 0.2); in contrast, the envelope-tracking response shown in the AEP was

[14] Electrocorticography (ECoG) is an invasive brain-imaging technique that records electrical activity of the brain directly from the surface of the cortex in surgical epilepsy patients.

deteriorated by the higher compression rates that reduced intelligibility, similar to Ahissar et al. (2001). These mixed results led to the conclusion that entrainment to the speech envelope in the auditory cortex is not necessarily a limiting factor for speech comprehension, which is consistent with what was suggested by Howard and Poeppel (2010).

As discussed previously, it is difficult to determine the causal relationship between speech comprehension and entrainment to the speech envelope, even if there is a positive correlation between the two. That is, it is not clear whether speech comprehension directly affects entrainment to the speech envelope, or entrainment purely reflects processing occurring early in auditory cortex which is merely modulated by acoustic properties of speech required for speech comprehension (e.g., spectral detail). In an attempt to resolve this issue, Millman et al. (2015) created a new paradigm in which subjects listened to identical tone-vocoded sentences before and after the original, unprocessed sentence was presented. After hearing the original sentence, perceptual “pop-out” is expected to occur (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005), where previously unintelligible vocoded speech becomes intelligible. Thus, the sentence presented before the original sentence served as a control for the sentence presented after the original sentence. In addition, they used tone-vocoded speech, as the amplitude envelope was shown to be better preserved in tone-vocoded speech than in noise-vocoded speech (Whitmal, Poissant, Freyman, & Helfer, 2007). They showed that phase-locked responses were not different between unintelligible and intelligible sentences that were acoustically

identical, suggesting that theta-band entrainment to the temporal envelope is not enhanced by linguistic information.

Taken together, the studies reviewed so far have yielded inconsistent results on the effect of speech intelligibility on cortical entrainment to the temporal envelope. Some of the inconsistencies may have arisen from using different speech intelligibility manipulations, or potential confounds associated with using some of the manipulations. For example, it has been suggested that the amplitude envelope may not be faithfully retained in noise-vocoded speech and time-reversed speech (e.g., Millman et al., 2015). In addition, it is sometimes difficult to determine the causal relationship between speech intelligibility and envelope tracking because any decrease in speech intelligibility may simply be an outcome of degraded processing of the acoustic signal at an auditory level (e.g., due to reduced spectral resolution), which would weaken the view of top-down amplification of envelope tracking (e.g., Peelle et al., 2013).

3.1.3 Cross-linguistic differences in neural entrainment to speech

In order to avoid potential confounds that arise from altering the acoustic signal, the current study manipulated speech intelligibility by comparing responses recorded from listeners with different language backgrounds who differ in terms of how much they can understand the target language. In addition to investigating the relationship between intelligibility and entrainment, the cross-linguistic experiment was able to examine the role of native language experience in cortical entrainment; greater entrainment for one's native language speech could be expected given that listeners'

perceptual representations are better tuned for their native language (e.g., Iverson et al., 2003). There is some neurophysiological evidence suggesting that language experience can alter how speech is processed even at early auditory levels. For example, the mismatch negativity (MMN) response (i.e., a pre-attentive brain response to an odd stimulus in a sequence of identical stimuli) has proven sensitive to the categorisation of phonemes (e.g., Näätänen et al., 1997; Dehaene-Lambertz 1997), and Chinese listeners have been shown to have a more robust frequency following response (FFR) in the auditory brainstem than English speakers in response to Mandarin tones (Krishnan et al. 2005, 2008, 2009; see Chapter 1.1 for details). While it has not yet been established how theta entrainment to the amplitude envelope can reflect cross-language differences, it is possible that listeners have facilitated neural synchronisation to certain shapes of amplitude envelopes that are related to the syllable structure (e.g., complex onsets) or rhythmic properties (e.g., stress) of their native language. Such cross-linguistic differences could be found at an auditory level without involving higher-level linguistic processing (i.e., during passive listening), as shown by other auditory brain responses.

Peña and Melloni (2012) had a cross-linguistic design in an EEG study where they had native Italian and Spanish speakers listen to Italian, Spanish, and Japanese sentences, with stimuli played both forwards and backwards. Using a time-frequency analysis, they examined the power of oscillations over time in each frequency band; theta (4-8 Hz), alpha (9-14 Hz) and middle gamma (55-75 Hz), rather than measuring the phase coupling between neural and acoustic signals. The results showed that listeners had a sustained increase in theta power when listening to all three languages. That is, theta-

band oscillatory activity occurred regardless of whether the language had similar rhythmic and prosodic structures to listeners' native language, or whether listeners understood the language. Peña and Melloni (2012) thus suggested that tracking of low-frequency amplitude fluctuations in the theta band occurs independently of higher-level language processing. In contrast, smaller theta power was found in backward-played versions of the stimuli; this occurred likely because syllables can be distorted in backward speech (e.g., ones containing plosives), thereby making it more difficult to track syllables.

In their study, cross-linguistic differences were only found in middle gamma power, which is thought to reflect semantic and syntactic unification processes (e.g., Hald, Bastiaansen, & Hagoort, 2006); listeners had enhanced middle gamma power only for the language that they were able to understand (i.e., their native language, forward). Although their findings suggest that theta oscillatory activity in response to speech is independent of comprehension and not language-specific, this remains to be confirmed largely because they examined theta-band power rather than phase entrainment. As mentioned previously, the phase modulation of theta oscillations is the key mechanism of envelope tracking (Luo & Poeppel, 2007; Howard & Poeppel, 2010).

3.1.4 Neural source localisation of the envelope tracking response

Generally, neural entrainment to the speech amplitude envelope has been shown to originate bilaterally in the auditory cortex, displaying similar spatial distribution as auditory M100 (e.g., Luo & Poeppel, 2007; Doelling et al., 2014). However, there is also evidence showing that entrainment to speech occurs in regions involved in higher-

level linguistic processing. Ding et al. (2016) examined cortical entrainment to speech using electrocorticography (ECoG), which provides better spatial resolution than MEG. They found significant syllabic-rate responses to intelligible speech (but not to the acoustic control) in bilateral posterior and anterior superior temporal gyri (pSTG and aSTG, respectively) and the left inferior frontal gyrus (IFG) when measured with high-gamma (70-200 Hz) power[15], and in broader areas in temporal and frontal lobes when measured with low-frequency activity. Some of these areas (e.g., left IFG, left pSTG, and right STG) are related to linguistic processing such as semantic, syntactic and prosodic processing (e.g., Hickok & Poeppel, 2007; Pallier, Devauchelle, & Dehaene, 2011). Similarly, Peelle et al. (2013) found bilateral phase-locked responses to the envelope of unintelligible speech in their MEG study in a number of regions including superior and middle temporal gyri, inferior frontal gyri, and motor cortex, but entrainment was enhanced for intelligible speech (i.e., 16-channel compared to 1-channel vocoded speech) in the left hemisphere, particularly around the left middle temporal gyrus. Because it is still a relatively low-level auditory area (e.g., Davis & Johnsrude, 2007), the authors concluded that this result supports top-down influences of linguistic content on low-level auditory processing.

Some studies have shown right hemisphere lateralisation in theta-band entrainment to the speech envelope (e.g., Luo & Poeppel, 2007), which supports the hypothesis that there is an inherent hemispheric asymmetry with the left hemisphere preferentially extracting information from short integration windows (20-50 ms) and the right hemisphere from long windows (150-250 ms; Asymmetric Sampling in Time;

[15] High gamma power is highly correlated with multiunit firing rates (e.g., Ray & Maunsell, 2011).

Poeppel, 2003). Gross et al. (2013) found the same lateralisation patterns; delta and theta entrainment were right-lateralised, whereas gamma entrainment was left-lateralised. That being said, most studies have failed to find right-lateralised theta-band entrainment.

3.1.5 Measures of cortical entrainment to the speech amplitude envelope

Because measuring phase-locking activity requires a high temporal resolution, most of the previous studies have used MEG. A few studies have used EEG (e.g., Hambrook & Tata, 2014; O'Sullivan et al., 2015) or ECoG (e.g., Nourski et al., 2009). There are also different ways of calculating the degree of phase-locking between neural oscillations and the amplitude envelope of speech. In the time-domain, one can measure the degree of cross-correlation between two signals (e.g., Ahissar et al., 2001; Nourski et al., 2009). Phase Locking Value (PLV) is also a metric of phase coupling which estimates the instantaneous phase difference between two signals (Lachaux, Rodriguez, Martinerie, & Varela, 1999; used in Gross et al., 2013). Computational modelling techniques such as signal reconstruction have also been widely used (e.g., Ding & Simon, 2012a; Di Liberto, O'Sullivan, & Lalor, 2015).

The present study used a measure of coherence, which is a metric that computes the degree of phase-locking between two signals as a function of frequency; the current study measured coherence between EEG signals and the amplitude envelope of the speech signals as used in Peelle et al. (2013), which is referred to as 'cerebro-acoustic coherence'. This measure was suitable for measuring neural processing for longer, continuous speech because it does not require repeated presentation of the same

stimuli. The present study was thus able to examine the processing of continuous speech that is more reflective of natural speech encountered in realistic environments.

3.1.6 Aims of the present study

It is a matter of continuing debate whether or not cortical entrainment to the amplitude envelope of speech is affected by higher-level linguistic processing. The aim of the present study was to see if entrainment to the temporal envelope of speech could be modulated by whether or not the listener understands the language. To avoid any confounds that may arise from altering the acoustic signal, an experiment was designed such that speech intelligibility (i.e., speech comprehension) could be manipulated using natural, unprocessed speech. To this end, a cross-linguistic study was conducted; EEG responses were recorded from two groups of listeners with different native language backgrounds - British English and Korean. They listened to continuous speech (i.e., stories) in three languages including their native language: English, Korean and Spanish. Similar to Peña and Melloni (2012), this cross-linguistic design created different conditions according to the degree of speech comprehension; English listeners could only understand English, but not Korean or Spanish. Korean listeners could understand Korean and English to some degree as they were second-language learners of English, but not Spanish.

In addition, the present study was conducted with the aim of developing EEG methods that measure how L2 listeners process continuous speech. By examining their cortical entrainment to the amplitude envelope of continuous speech, the present study may be able to reveal L2 speech processing difficulties for natural, continuous speech that

stem from an early auditory level. Specifically, native listeners' envelope tracking activity could be stronger than that of non-native listeners if linguistic processing (i.e., speech comprehension) can enhance lower-level tracking of the speech envelope. Native listeners can be expected to have more robust entrainment also because they can have perceptual representations that are tuned for the language (e.g., syllable structure, rhythm; see Chapter 3.1.3 for details).

The EEG experiment of the present study consisted of passive and active listening tasks. While most previous studies have used active listening tasks (e.g., comprehension tasks) to measure phase-locked responses to speech, a passive listening task was conducted first. Because auditory cortical responses can be recorded passively (e.g., Cortical Auditory Evoked Potentials), this task allowed for measurement of envelope tracking while listeners passively listened to speech without focusing attention on the linguistic content of the speech (i.e., they were watching a silent movie). Secondly, the active listening task was conducted using the same stimuli (different subjects). In this task, to ensure that subjects attend to the acoustic stimuli, listeners were asked to perform a syllable-spotting task. Listeners' attention to the stimuli was thus better controlled compared to the passive task. This particular task was chosen instead of speech comprehension tasks because listeners had to listen to languages that they did not understand. Meanwhile, listeners were naturally able to understand the speech materials while doing the syllable-spotting task, when the stimuli were spoken in their native language (or second language).

3.2 Methods

3.2.1 Subjects

Twelve monolingual native speakers of Standard Southern British English (6 females; mean age: 32.8 yr) and twelve monolingual native speakers of Korean (7 females; mean age: 28.4 yr) participated in the passive listening task. For the active listening task, there were also twelve monolingual native speakers of Standard Southern British English (8 females; mean age: 23.7 yr) and twelve monolingual native speakers of Korean (7 females; mean age: 25.6 yr). The groups in the passive and active tasks were independent of each other (i.e., no subject took part in both tasks). The listening task was a between-subjects factor because this study was initially designed as a passive listening task, but the active task was conducted afterwards to see if focusing attention on the acoustic stimuli affects envelope-tracking activity (see Chapter 3.1.6). There were four more Korean speakers (4 females) who participated in the experiments, but their data were not used for analysis due to containing several noisy electrodes or an excessive amount of blinking. All subjects were right-handed with no self-reported hearing, language or neurological disorders.

All English subjects reported that they had never learned Korean. All Korean participants reported that they had learned English for 12.6 years on average starting from approximately the age of 12, and that they had lived in English-speaking countries for an average of 14 months as adults; they were living in London at the time of testing. Most of the English and Korean subjects had not learned Spanish. However, one English subject in the passive listening task reported that he was learning Spanish at the time of testing, and that his proficiency level was beginner. Three English

subjects in the active task also reported that they had learned Spanish at school for approximately 3.7 years on average, but their self-reported proficiency level was beginner or intermediate. Some of them also informally reported that they can only understand basic Spanish words and expressions. Two Korean subjects in the active listening task reported that they had learned Spanish in Korea at the age of 19 and 16. However, their experience with Spanish was limited; their length of learning was one year and their self-reported proficiency level was beginner or intermediate. Considering that the Spanish stories used in the experiment were classic novels read by a native speaker at a natural speaking rate, it seems unlikely that these subjects with limited experience with Spanish were able to understand the stories. Nonetheless, additional analyses were carried out to see if the experience with Spanish affected their neural entrainment.

3.2.2 Stimuli

Female speakers of the three languages (i.e., English, Korean, and Spanish; one speaker each) were recorded reading stories in their native language. Specifically, the English talker was a native speaker of Standard Southern British English (age: 25), the Korean talker (i.e., the author) was a native Korean speaker from a city near Seoul (age: 28), and the Spanish talker was a native Spanish speaker from Galicia (i.e., north west Spain; age: 25). Stories were excerpted from two or three short stories or novels for each language; the Secret Garden (Burnett, 1909) and Lazy Jack (n.d.) for English; Heungbuwa Nolbu (2003) and Mongsil Unni (Kwon, 2007) for Korean; and Casa tomada (Cortázar, 1947), La casa de Asterión (Borges, 1949), and El eclipse (Monterroso, 1958) for Spanish. The recordings were edited such that each stimulus

comprised two minutes of continuous speech excerpted from a single story. Six stimuli were created for each language (6*3=18). The same 18 stimuli were used for the passive and active listening tasks. All stimuli had 44100 16-bit samples per second. The RMS amplitude was equalised to 70 dB SPL across stimuli.

In the active listening task, subjects were asked to count how many times a target syllable occurred in each stimulus. Because of large dissimilarities between phoneme inventories of the three languages, it was not possible to choose the same target syllables for all three languages. Different target syllables were therefore used in each language as shown in Table 3-2. However, they were carefully selected after consulting with native speakers of each language so that each target syllable consisted of a consonant and a vowel that are not difficult to detect acoustically for both native and non-native listeners of the language. The target syllables occurred 1 to 6 times in each 2-minute stimulus.

Table 3-2: Target syllables used for the syllable-spotting task

English	Korean	Spanish
[kæ]	[sɛ]	[na]
[dɪ]	[pa]	[tʃe]
[hɛ]	[hɛ]	[xo]
[bɑː]	[sa]	[tʃo]
[mæ]	[wa]	[xi]

3.2.3 Apparatus

All stimuli were presented via Praat (Boersma & Weenink, 2014) using an external sound card (RME Fireface UC) and Etymotic ER-1 insert earphones. To obtain timing

information of the stimuli, triggers were generated as pulses on a separate audio channel, which were converted to TTL triggers via a custom circuit. EEG was recorded through a Biosemi Active Two system with 64 (Ag/AgCl) electrodes mounted on an elastic cap, and 7 external electrodes (left and right mastoids, nose, two vertical and two horizontal EOG electrodes). Unreferenced EEG signals were recorded with a sampling rate of 2048 Hz. Electrode impedances were kept within the range of ±25k Ω during the experiment. Time-aligned triggers were also recorded by the EEG system.

3.2.4 Procedure

In the passive listening task, subjects were instructed to watch a silent animation movie and not to pay attention to the speech being presented through insert earphones, while their electrophysiological activity was recorded. The stimuli were presented in six randomized blocks, each consisting of three stimuli (one language each), and the subjects had a short break between blocks.

In the active listening task, EEG was recorded while subjects fully attended to the stimuli, which were also presented in six randomized blocks. Before subjects began listening to each stimulus, they listened to an instruction in their native language recorded by the same speaker who read the stimuli of that language, to say that they should count how many times the target syllable occurs in the two-minute story. The target syllable was presented in isolation three times in a row during the instruction. They were also informed that they could only use their fingers for counting if needed, in order to minimize any body movements. After each stimulus finished, the subjects

orally answered how many times they heard the target syllable. They had a short break between blocks. Because the sole purpose of conducting the syllable-spotting task was to make sure that subjects paid attention to the stories, the results of this task were not analysed.

3.2.5 Analysis

3.2.5.1 Pre-processing

All pre-processing of the EEG data was performed offline in Matlab. The EEG signals were referenced to the average of the left and right mastoids. Noisy channels were interpolated. The data were then high-pass filtered at 0.1 Hz and low-pass filtered at 40 Hz using Butterworth filters as implemented in the ERPlab toolbox (Lopez-Calderon & Luck, 2014) of EEGlab (Delorme & Makeig, 2004). All pre-processing procedures, except for filtering, were performed in Matlab using the Fieldtrip toolbox (Oostenveld, Fries, Maris, & Schoffelen, 2011).

3.2.5.2 Coherence analysis

In the present study, coherence was used to measure the degree of phase-locking between the amplitude envelope of speech and its corresponding EEG signal (i.e., cerebro-acoustic coherence, Peelle et al., 2013). Prior to calculating coherence, amplitude envelopes were calculated from the speech stimuli; the speech signals were full-wave rectified and then filtered using the same high-pass (cut-off: 0.1 Hz) and low-pass filters (cut-off: 40 Hz) that were used for the EEG signals (i.e., Butterworth filters in ERPlab). The speech signals were also down-sampled to 2048 Hz to match

the same sampling rate of the EEG data. These procedures were all performed in Matlab.

The continuous EEG signal and the amplitude envelope of its corresponding acoustic signal were segmented into 2-second epochs. Because there were some silent portions in the stimuli (e.g., pauses between sentences), epochs were rejected if the RMS amplitude of the corresponding speech signals was smaller than the lowest 10^{th} percentile of the RMS amplitude for the entire speech data. The epochs were then multiplied by a Hanning-window and transformed into the frequency domain using a Fast Fourier transform (FFT). The 2-second signals used in the Fourier transform resulted in a 0.5 Hz frequency resolution. As shown in the following formula (3.1), coherence between two signals x and y is defined as the cross-spectral density of the two signals G_{xy} divided by the power spectrum of each signal $G_{xx}$and G_{yy}, and each of these components is averaged across trials (i.e., epochs) before calculating coherence. Values of coherence lie between 0 and 1 in which 0 means no phase coupling and 1 means perfect phase coupling; the more constant the phase difference between the two signals is, the closer the coherence value is to 1.

$$C_{xy} = \frac{|{G_{xy}}^2|}{G_{xx}G_{yy}} \tag{3.1}$$

3.2.5.3 Denoising Source Separation

Denoising Source Separation (DSS; de Cheveigné & Simon, 2008) was used to isolate the neural activity that was phase-locked to the amplitude envelopes of the stimuli. This technique increases the signal-to-noise ratio of the activity of interest within

neural data by deriving linear combinations of electrodes with weights. Specifically, DSS is performed by first running a Principle Component Analysis (PCA) on raw neural data, normalising it (i.e., rendering the data 'spherical'), and then applying 'bias filters' such as data averaged across trials to enhance relevant parts of the data (i.e., the power along the direction that captures that activity of interest) while reducing the rest. PCA is then applied again to align these directions with the final component axes (such that the highest-ranked component captures the relevant activity best). This technique can be used to remove artifacts (e.g., power line noise, cardiac artifacts) or to extract part of data that is only relevant to a specific response (e.g., auditory-specific response during audiovisual speech perception; de Cheveigné & Simon, 2008; de Cheveigné & Parra, 2014). DSS has also been shown to be effective in isolating the speech-tracking response in several previous studies (Ding & Simon 2012a; Ding & Simon 2014; Kong, Somarowthu, & Ding, 2015; Ding et al., 2016).

In the current study, spatial filters (i.e., linear combinations of electrodes) were calculated for each subject from the covariance of the raw data at each electrode and the covariance of the coherence values at each electrode averaged over all trials between 1 and 20 Hz. The present study used the first four DSS components that maximized the reliability of coherence for each subject. It appeared that the first four components all captured activity related to envelope tracking, but with varying degrees. The components were then projected back into sensor space and were used to calculate coherence.

3.3 Results

As displayed in Figure 3-1, listeners had clear coherence peaks in the theta range (4-8 Hz) across conditions, suggesting that listeners had envelope-tracking activity in this frequency range consistent with the previous literature. Figure 3-2 shows the topographic distribution of the mean coherence values for each group of listeners (i.e., 2 listening tasks * 2 listener native languages). The topographies suggest that the envelope-tracking response was strong in frontocentral electrodes. While this should be interpreted with caution without appropriate source analyses, the distribution broadly agrees with other studies that found speech entrainment or other auditory responses in similar areas (e.g., Luck & Kappenman, 2012; Doelling et al., 2014). Coherence values were therefore averaged across 25 frontocentral electrodes (Fz, F1, F2, F3, F4, F5, F6, F7, F8, FCz, FC1, FC2, FC3, FC4, FC5, FC6, FT7, FT8, Cz, C1, C2, C3, C4, C5, C6) for statistical analysis. Other electrode sites were not included in the main statistical analysis because the data that were projected back into sensor space from DSS components were already weighted, such that electrodes that showed great phase-locked responses to the speech envelope had greater weights than electrodes that did not. Comparing different electrodes in the statistical analysis is thus unlikely to yield different results.

Figure 3-1: Results of the coherence analysis by language for all listening tasks and listener native languages (i.e., for each group of listeners). Coherence values are plotted as a function of frequency (2-20 Hz).

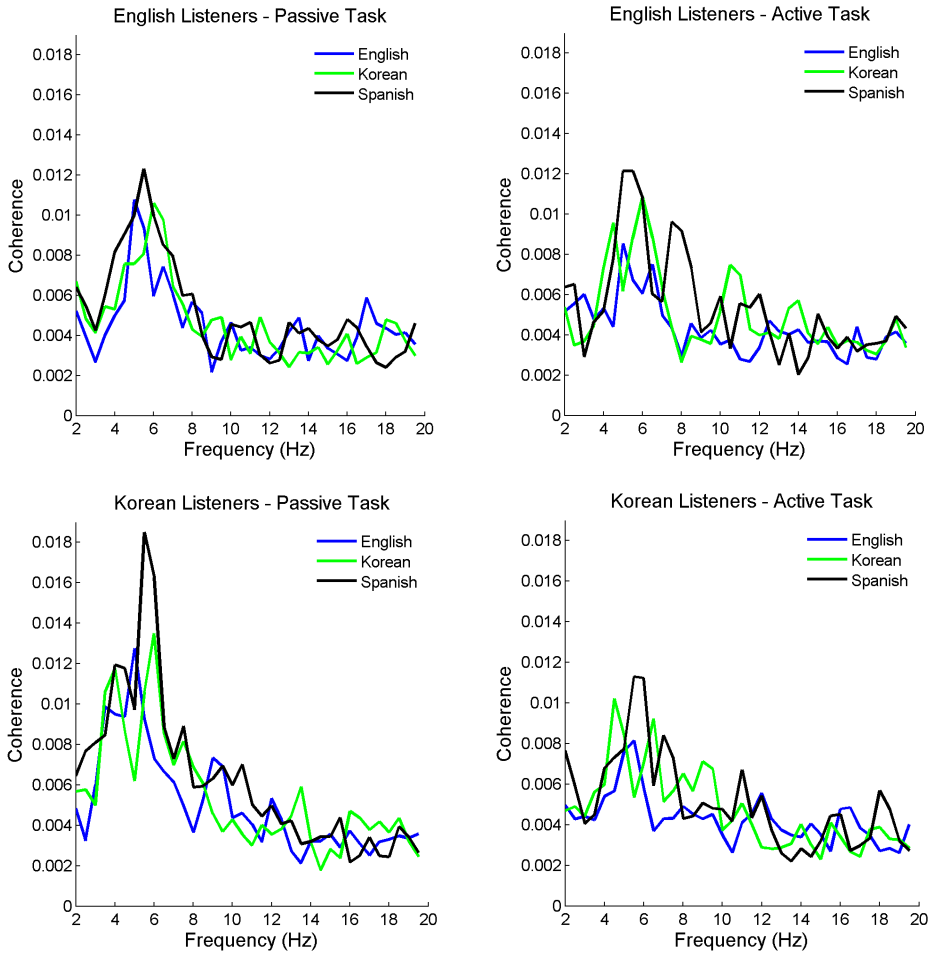
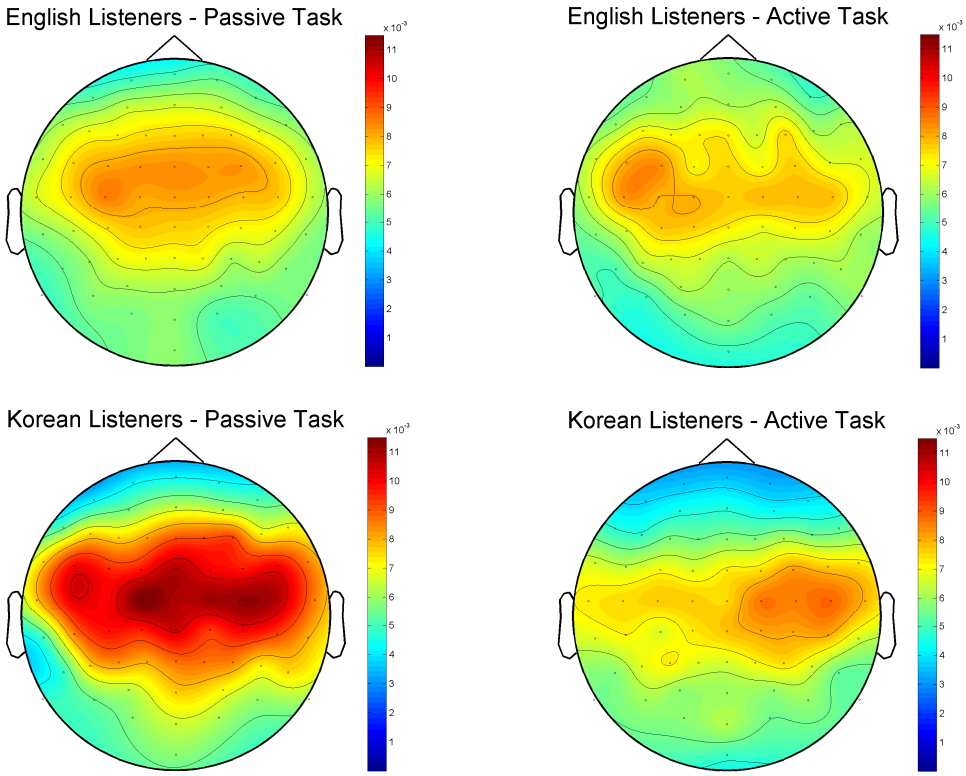


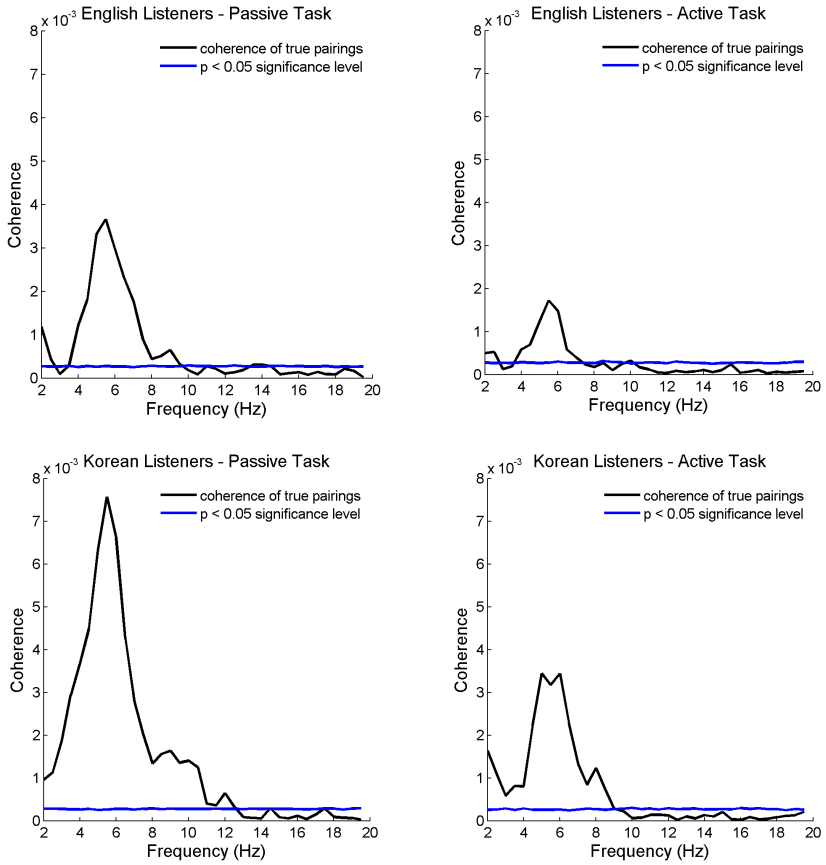
Figure 3-2: Topographies of the mean coherence values in the theta range (4-8 Hz) for all listening tasks and listener native languages (i.e., for each group of listeners)



Prior to conducting the main statistical analysis, a permutation analysis was conducted; within each group of listeners, amplitude envelopes were randomly permuted between trials to calculate the coherence for random pairs of amplitude envelopes and EEG signals. This process was performed 500 times to determine the distribution of the permuted coherence. The coherence was again averaged across the 25 frontocentral electrodes in this analysis. Any real coherence values above the 95th percentile of the permuted coherence were considered to be significant. As displayed in Figure 3-3, coherence values of the real data (i.e., true pairings of EEG and amplitude envelopes) lie above the significance line in the theta range (4-8 Hz) in all conditions. That is, the results confirm that the coherence peaks observed in the current study were indeed driven by greater phase coupling between EEG and speech signals occurring in that

range. It should be noted that in this permutation analysis, the coherence calculation shown in (3.1) was performed across all subjects within each group (rather than calculating the coherence within each individual and averaging it across individuals), under the assumption that all individuals within each group would have more or less the same phase relationships with the amplitude envelopes. However, the overall magnitude of coherence was found to be smaller in this analysis, because the calculation performed across subjects was in fact affected by the inter-individual variation of the data especially in the active listening task.

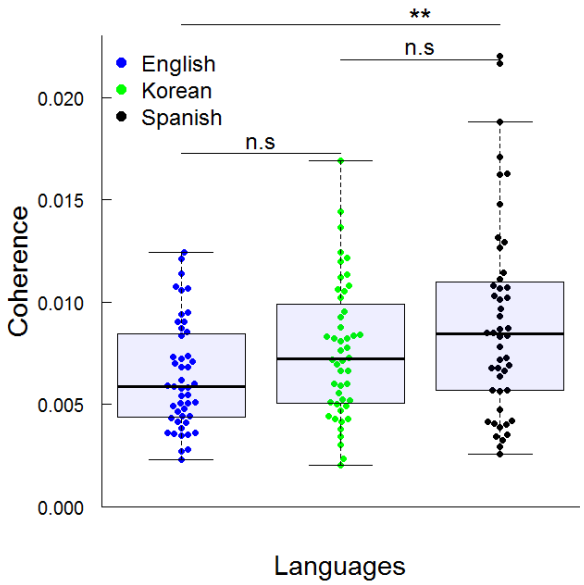
Figure 3-3: Permutation analysis for all listening tasks and listener native languages (i.e., for each group of listeners). Black lines show mean coherence values calculated across subjects as a function of frequency. Blue lines denote the p < 0.05 significance level based on the distribution of the permuted coherence (500 random permutations).



Based on the results of the permutation analysis, the main statistical analysis was performed with coherence values averaged between 4 and 8 Hz (9 frequency points at intervals of 0.5 Hz). A linear mixed-model analysis was conducted in R using the R package lme4 (Bates et al., 2015), with coherence values as the dependent variable; language background of listeners (English and Korean listeners), language of stimuli (English, Korean, and Spanish), and listening task (passive and active) as fixed effects; and with by-subject random intercepts. Specifically, all the main effects of the fixed factors and their interactions were included in the model, as it was expected that they all might affect the degree of neural entrainment to the amplitude envelope (see Chapter 3.1.6). The package CAR (Fox & Weisberg, 2002) was used to calculate type II analysis-of-variance tables.

There was a significant main effect of language of stimuli, $\chi^2(2) = 26.55$, $p < 0.001$. Bonferroni post-hoc t-tests found that the coherence for Spanish was significantly greater than for English, $p = 0.0017$, while the difference between Spanish and Korean and the difference between Korean and English were not significant, $p = 0.1818$ and 0.3162, respectively (Figure 3-4). However, the interaction between listener's native language background and stimuli language was not significant, $p = 0.9353$. That is, the effect of stimuli language (i.e., greater entrainment for Spanish) was not different between English and Korean listeners.

Figure 3-4: Combined boxplot and beeswarm plot of individual coherence values for each language of stimuli averaged over English and Korean listeners.



One could expect that the interaction between listener and stimuli language would only emerge in the active listening task, given the previously found positive relationship between speech comprehension and entrainment (e.g., Peelle et al., 2013). However, there was no significant 3-way interaction between stimuli language, listener's native language and listening task, $p = 0.5791$. This suggests that entrainment to the speech envelope is not modulated by whether or not listeners understand the language of the speech materials (i.e., active task), or whether or not they process rhythmic or syllabic structures of their native language. The main effect of listening task also did not reach significance with the p-value of 0.1103, indicating that neural oscillations can become phase-locked to speech in the environment without listeners' paying attention to it. All

the other main effects and interactions were also not significant[16], $p > 0.05$. Coherence results for each stimulus language, listener native language and listening task are displayed in Figure 3-5 and Table 3-3.

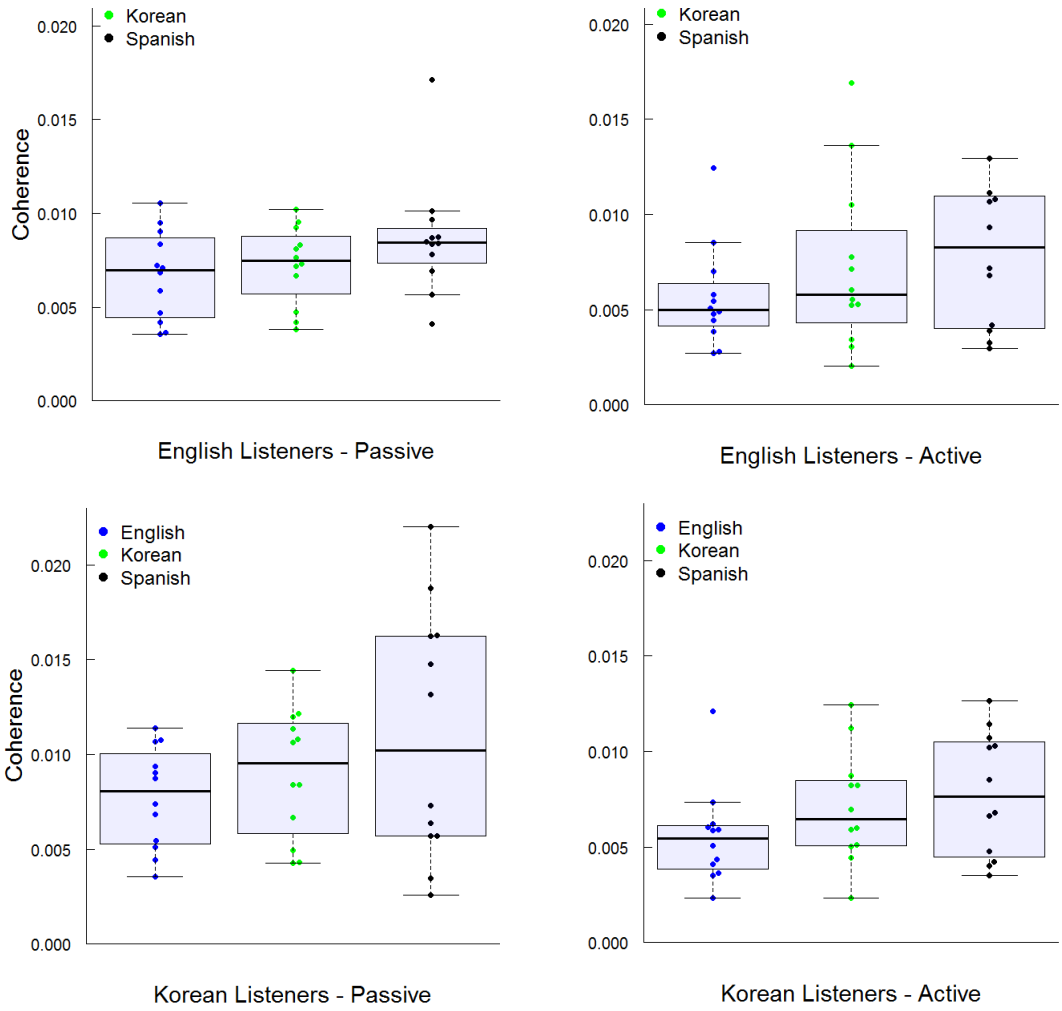
To examine whether the results were different for the six subjects who had learned Spanish (see Chapter 3.2.1), additional analyses were performed. A linear mixed-model analysis was conducted with coherence values as the dependent variable; listeners' experience with Spanish (no experience vs. experience) and language of stimuli as fixed effects; and with by-subject random intercepts. The main effect of language was also significant in this analysis, $\chi^2(2) = 27.42$, $p < 0.001$. However, the main effect of Spanish learning experience failed to reach significance with the p-value of 0.0783. More importantly, the two-way interaction between stimuli language and Spanish experience was not significant, $p = 0.7365$, confirming that the subjects' experience with Spanish did not affect their neural entrainment to that language. Furthermore, the results of the main mixed-model analysis were highly similar when it was conducted without the six subjects. That is, the main effect of stimuli language was significant, $\chi^2(2) = 22.27$, $p < 0.001$, with Spanish showing higher coherence than English (Bonferroni post-hoc t-tests, $p = 0.0041$), but no other main effects or interactions were significant.

[16] However, it is unclear why Korean listeners had a stronger envelope-tracking response in the right hemisphere than did English listeners particularly in the active listening task (Figure 3-2). It remains for further research to address this question using appropriate source localisation methods.

Table 3-3: Descriptive statistics of coherence for each group of listeners (i.e., 2 listening tasks * 2 listener native languages), separately for different language conditions

Listener Group	Language	Mean	Standard Deviation
English listeners Passive task	English	0.0067	0.0024
	Korean	0.0072	0.0021
	Spanish	0.0087	0.0031
English listeners Active task	English	0.0056	0.0027
	Korean	0.0072	0.0044
	Spanish	0.0087	0.0053
Korean listeners Passive task	English	0.0077	0.0027
	Korean	0.0090	0.0034
	Spanish	0.0110	0.0066
Korean listeners Active task	English	0.0055	0.0025
	Korean	0.0071	0.0029
	Spanish	0.0078	0.0032

Figure 3-5: Combined boxplot and beeswarm plots of individual coherence values by language for all listening tasks and listener native languages (i.e., for each group of listeners)



3.4 Discussion

The present study investigated neural processing of continuous speech in EEG recordings in terms of cortical entrainment to the amplitude envelope of speech. A cross-linguistic paradigm was used to vary the degree of speech comprehension/intelligibility without changing the acoustic properties of the speech. The present study reproduced the previous finding (e.g., Luo and Poeppel, 2007; Peelle et al., 2013) that brain oscillations become phase-locked to the amplitude envelope of the speech signal in the 4-8 Hz range (i.e., theta), with a broad maximum at anterior and central electrodes. However, the results demonstrate that neural entrainment to the speech envelope is not modulated by whether or not listeners understand the speech in contrast to what some of the previous studies found (see Table 3-1). It was only modulated by the language of stimuli; both English and Korean listeners had higher entrainment to Spanish, the language that they both could not understand.

It is apparent that entrainment to Spanish stimuli was not enhanced by speech comprehension, because none of the listeners were able to understand Spanish. Furthermore, the effect of language was found across different listening tasks. This suggests that the coherence difference between languages was purely attributable to acoustic properties of the speech signals such as speaking rate or rhythm characteristics, rather than other higher-order effects such as attention or comprehension. One could argue that this result was driven by differences in speech rhythm between the languages. Specifically, syllables are produced at relatively more regular intervals in Spanish (i.e., closer to syllable-timing) compared to stress-timed languages such as English where the duration of syllables varies to a larger extent due

to having stressed and unstressed syllables (e.g., Ramus, Nespor, & Mehler, 1999). The rhythmic classification of Korean is less clear because it has features of both syllable-timed and stress-timed rhythm (e.g., Lee et al., 1994; Seong, 1995). One may thus think that the relatively regular occurrence of syllables in Spanish facilitated neural phase locking to the temporal envelope (e.g., by requiring less frequent phase adjustment). However, this result should be interpreted with caution because the speech materials were recorded by one speaker for each language. That is, other individual speaker characteristics may have contributed to the differences in coherence such as speaking rate.

Regardless of what acoustic properties of the Spanish stimuli gave rise to the increase in entrainment, the present findings suggest that the reduction in neural phase-locking that had been previously seen for less intelligible speech may have been caused by reduced spectral modulations (e.g., vocoded speech) or other changes in the speech signal (e.g., altered amplitude envelope), rather than resulting directly from listeners' inability to use linguistic knowledge. This is consistent with the collective feature tracking hypothesis (Ding & Simon, 2012b): entrainment to the amplitude envelope may be an index of collective neural encoding of all speech features rather than the envelope on its own. That is, it appears that envelope tracking occurs purely at an auditory level during speech processing, independently of higher-level linguistic processes.

Researchers have also theorised that phase entrainment to the temporal envelope can measure tracking of syllables (Giraud & Poeppel, 2012; Ghitza, 2013). Similarly, it

was also expected in the current study that listeners might have stronger entrainment for their L1 speech because they can have perceptual representations that are better tuned for their L1 rhythm or syllable structure through language experience, regardless of speech comprehension. However, the effect of language experience was not found in low-frequency entrainment to the speech envelope in this study (see further discussion in Chapter 5). Similarly, a recent study by Ding and colleagues (2016) found that both native speakers of English and Mandarin had entrainment to Mandarin sentences at syllabic rate (i.e., theta), whereas entrainment to larger linguistic structures – sentences and phrases in the delta range (i.e., around 1 and 2 Hz, respectively) was only found in native Mandarin listeners, who could understand the sentences and parse them into appropriate linguistic units. This multi-timescale entrainment to speech may be a neural mechanism that is specific to speech processing (Giraud & Poeppel, 2012), but it appears that neural entrainment to rhythmic sensory inputs (e.g., syllables) itself is not; it has been found in non-human primates such as macaque monkeys for both speech and non-speech sounds such as animal vocalisations (e.g., Lalor, Power, Reilly, & Foxe, 2009; Steinschneider, Nourski, & Fishman, 2013).

There was also no difference in the observed coherence between passive and active listening tasks, suggesting that theta entrainment to the acoustic envelope of speech is observed even when listeners are not paying attention to the speech. One could argue that this result is not consistent with some of the previous studies that found significant effects of attention on speech entrainment (e.g., Kerlin et al., 2010; Ding & Simon, 2012a). However, these effects have only been found in complex auditory scenes;

attention has been shown to selectively enhance the neural entrainment to the target talker over the distractor. Moreover, low-level auditory brain areas have been found to maintain representations of both attended and unattended speech signals (Ding & Simon, 2012a; Zion Golumbic et al., 2013). It has also been suggested that increased neural phase-locking to target talkers is related to successful segregation and encoding of the target speech signals (Ding & Simon, 2012a). In the current task, entrainment was similar in passive and active listening tasks, likely because listeners were able to process the speech signal without needing to segregate competing speech streams with greater attention as required in competing-talker situations. Thus, this finding does not necessarily contradict the earlier studies.

In summary, the present study was able to observe phase-locked neural responses to continuous speech signals in EEG recordings by measuring the degree of phase-locking between neural oscillations and the speech envelope using 'coherence', which had been previously used in Peelle et al. (2013) with MEG data. It was also possible to extract envelope-tracking components using DSS (de Cheveigné & Simon, 2008), which allowed clearer observation of the activity. This study was different from most of the previous studies in that it compared neural entrainment between native English and Korean listeners when they were listening to different languages, without manipulating the acoustic properties of the signals. The present work brings a better understanding of the link between speech comprehension and cortical entrainment to the speech envelope. It appears that envelope tracking occurs mainly at an early auditory level of speech processing and is separable from higher levels of linguistic processing. It seems to be purely modulated by acoustic properties of the speech signal

rather than listeners' higher-level linguistic knowledge. The current findings therefore suggest that this neural measure may not be suitable for investigating L2 speech recognition difficulties that are directly caused by less-developed perceptual and linguistic representations of L2 listeners.

Chapter 4 Speech recognition in multi-speaker environments[17]

4.1 Introduction

In everyday life, listeners often encounter situations in which multiple talkers are speaking at the same time as in social gatherings like parties. Concurrent speech signals can physically mask the target speech signal, rendering it less intelligible. Moreover, selectively attending to one speech signal and tuning out other signals places additional demands on attention and cognitive control (e.g., Brungart, 2001). Given that following a conversation in a second language can be effortful even in quiet, the cognitive and perceptual demands of listening in multi-talker environments can make L2 speech recognition doubly hard and lead to more recognition errors (Cooke et al., 2008). In addition, as shown in Study 1, speech recognition in adverse conditions can be modulated by the accents of the listener and the talker (e.g., Bent & Bradlow, 2003). While individuals (i.e., especially native listeners) can comprehend accented speech with ease in quiet listening conditions, dealing with pronunciations that deviate from the listener's own norms can be hard in a competing-talker environment.

This chapter details a study which investigated speech processing by L1 and L2 listeners in a two-talker situation where they selectively attended to the target talker over the distracting talker (i.e., presented dichotically). Furthermore, the accent of the talker varied such that it either matched that of the listener or not. This study used EEG methods to tap into specific speech recognition processes rather than assessing overall

[17] Part of this work has been submitted to a journal as: Song, J., and Iverson, P. (under revision). Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents.

speech recognition performance alone; measures of neural entrainment and N400 were used as well as a behavioural measure of speech recognition, to more comprehensively investigate how L1 and L2 listeners modulate their auditory and lexical processing in these difficult listening conditions.

Although the results of Study 2 suggested that neural entrainment to speech is not sensitive to listeners' native language experience or linguistic processing, one could still expect that native listeners have stronger entrainment to the target speech than do non-native listeners in a two-talker situation, because greater target-talker entrainment has been found for more intelligible speech in previous studies, due to better speech segregation or comprehension (Kong et al., 2015; Rimmele, Zion Golumbic, Schröger, & Poeppel, 2015). However, target-talker entrainment in a competing-talker environment can be enhanced by top-down attention (e.g., Ding & Simon, 2012a; Kerlin et al., 2010), which can increase as speech recognition becomes more difficult. It is thus possible that speech entrainment by native and non-native listeners is modulated by a more complex relationship between intelligibility and their recognition difficulty. The N400 response (Kutas & Hillyard, 1980) was used to measure neural effort that listeners exert for lexical processing in given semantic contexts; lexical processing can be hindered in adverse conditions, but listeners can also increase the degree of lexical processing to help overcome difficult listening conditions (e.g., accent; e.g., Romero-Rivas, Martin, & Costa, 2015).

4.1.1 Speech recognition in multi-speaker environments

In everyday life, listeners often encounter a situation where they need to selectively focus on the speech of one talker that is masked by the speech from other talkers. Speech segregation can be performed using spatial cues if sound sources come from different locations, or acoustic-phonetic differences between target and distracting speakers such as their vocal characteristics, intensity levels, or prosodic features (e.g., Darwin & Hukin, 2000). Even if the target and distracting signals can be successfully segregated, the presence of competing speech can cause two types of masking (e.g., see Brungart, 2001; Mattys et al., 2009 for reviews). Energetic masking occurs when the target speech signal is physically masked by another signal in shared spectro-temporal regions. As a result, portions of the target speech signal are rendered inaudible. When this occurs, listeners can use "glimpses" (i.e., spectro-temporal regions where the target speech signal is least affected by the distractor) to help identify the speech signal (Cooke, 2006)

The other type of masking is called 'informational masking', which refers to any masking effects that remain once energetic masking has been accounted for (Cooke et al., 2008). According to Cooke et al. (2008), this mostly refers to higher-level (i.e., cognitive or linguistic) consequences of masking. Specifically, listeners need to exert additional cognitive resources to ignore the distracting signal (i.e., competing attention of the masker). If the competing speech is intelligible, there is further interference from the linguistic (i.e., lexical and semantic) content of the competing signal. Competing speech is thus more detrimental to the recognition of the target speech when it is produced in a native language than in a foreign language (Rhebergen et al., 2005; Van

Engen & Bradlow, 2007; Garcia Lecumberri & Cooke, 2006). Similarly, babble noise is more detrimental when it is produced by few talkers than by a number of talkers (e.g., Freyman et al., 2004). That is, because any processing caused by the masker (e.g., competing attention and linguistic interference) can tax cognitive resources, the performance on the target speech recognition can suffer from the overall increase in cognitive load. Furthermore, listeners can misallocate elements of the masker (e.g., frication, bursts, or larger units such as words) to the target, thereby causing an error in the identification of phonemes and words in the target signal. This is also categorised as one type of informational masking (Cooke et al., 2008).

4.1.2 Increased cognitive load in adverse listening conditions

In the present study, listeners selectively attended to the target talker presented in one ear while ignoring the distracting talker in the other ear. As a result, informational masking mainly occurred. That is, the listener could hear the target speech signal clearly in the target ear (i.e., no energetic masking), but they needed greater cognitive resources or listening effort to process the target speech. The same holds true for other realistic communicative situations such as dual-tasking (e.g., pilots who need to control the flight of an aircraft while communicating with air-traffic controllers). Because one's processing resources are limited (Kahneman, 1973), listeners under these circumstances have less cognitive resources left to allocate to the main speech recognition task.

Previous research has suggested that the accuracy of speech perception generally suffers from the cognitive load caused by concurrent tasks (e.g., visual search), but it can also alter specific aspects of speech perception. Specifically, listeners showed an

inflated “Ganong effect” for phoneme categorisation; an acoustically ambiguous phoneme is more likely to be categorised as what results in a word (e.g., “gift” rather than “kift”), but this tendency was greater while listeners were performing a concurrent task (Mattys & Wiget, 2011). That is, listeners’ ability to pay attention to fine phonetic detail is disrupted under high cognitive load. Similarly, a concurrent visual task increased listeners’ reliance on lexical cues to word segmentation while reducing reliance on acoustic cues (Mattys et al., 2009). Mattys et al. (2014) has suggested that cognitive load in fact disrupts perceptual sensitivity only at the sub-lexical level without affecting lexical activation; they found similar lexical effects on phoneme restoration in load and no-load conditions. Likewise, cognitive load caused by a visual task was shown to reduce auditory responses to unattended non-speech tones that were presented concurrently with the visual stimuli, causing “attentional deafness” (Molloy, Griffiths, Chait, & Lavie, 2015). In this series of studies, cognitive load refers to any load that arises from the recruitment of central processing resources due to simultaneous attentional and mnemonic processing that is not related to speech recognition (Mattys & Wiget, 2011).

More generally, additional listening effort or cognitive load can arise as a secondary consequence of adverse listening conditions, whether due to background noise, accented speech, or hearing loss. The Ease of Language Understanding (ELU) model (e.g., Rönnberg, 2003; Rönnberg, Rudner, Foo, & Lunner, 2008) suggests that explicit working memory processes are needed to resolve mismatches between listeners’ phonological representations and the acoustic input, resulting in increased cognitive effort. The increased demands on working memory can lead to a decrease in

performance on other tasks due to a shortage of cognitive resources. For example, adults with hearing loss were found to be poorer at remembering sentences that they heard than those with normal hearing (e.g., Piquado, Benichov, Brownell, & Wingfield, 2012).

Various methodologies have been used to investigate listening effort. Neuroimaging studies have demonstrated that listeners recruit additional cognitive resources in adverse conditions, by showing greater activation in middle and superior temporal areas, left inferior frontal gyrus, premotor cortex, and other brain areas that are not typically engaged during language processing, including anterior insula, frontal operculum, and anterior cingulate (e.g., Davis & Johnsrude, 2003; Eckert et al., 2009; Erb, Henry, Eisner, & Obleser, 2013; Erb & Obleser, 2013; Vaden et al., 2013; Wild et al., 2012). This increased effort can sometimes play a compensatory role. For example, engagement of these additional areas (e.g., engagement of middle temporal gyrus by older listeners) was shown to improve speech comprehension (i.e., behavioural performance) in some studies (e.g., Erb & Obleser, 2013; Peelle et al., 2011; Vaden et al., 2013).

In contrast to those functional MRI studies, EEG can provide a means of investigating listening effort that is more temporally precise. For instance, the power of alpha oscillations has been shown to increase when processing more degraded auditory stimuli, and the N1 response (i.e., a negative peak occurring at around 100 ms after stimulus onset) was likewise found to increase and peak earlier (e.g., Obleser, Wöstmann, Hellbernd, Wilsch, & Maess, 2012; Obleser & Kotz, 2011). Recently,

physiological measures such as pupillometry (i.e., measure of pupil dilation) have also been used. For example, Zekveld, Kramer, and Festen (2011) found that listening effort as measured by the pupil response declined as the intelligibility of speech increased, but this decrease in the pupil response was smaller for older and hearing-impaired listeners than for normal-hearing listeners. That is, older and hearing-impaired listeners exhibited less 'release from effort' as a function of intelligibility. One can also use dual-tasking paradigms to assess listening effort, because a decrease in performance on a secondary task can reflect an increase in listening effort associated with the primary speech perception task (see McGarrigle et al., 2014 for a review).

4.1.3 Listening effort during L2 speech recognition

The main focus of this study was on how L2 speech processing is affected by increased speech recognition difficulties in a competing-talker environment. It is well-established that adverse listening conditions are more detrimental to L2 speech perception (see Chapter 1.3 for details), but the previous findings were largely based on the effects of energetic masking (e.g., stationary noise). That is, there is only sparse evidence for effects of cognitive load on L2 speech perception, and previous results are not congruent with one another.

In a visual world eye-tracking study by Ito, Corley, and Pickering (2017), cognitive load caused by a concurrent memory task delayed predictable eye movements using preceding verbs, but to similar degrees for L1 and L2 listeners. However, Cooke et al. (2008) found that non-native listeners were more adversely affected by informational masking caused by competing speech streams, and that the native advantage became

greater with increasing levels of the masker. The authors suggested that this occurred likely because non-native listeners were less accurate at allocating sound components to the target/masker using language-specific cues (e.g., accent-related acoustic-phonetic characteristics of the competing talkers) and they suffered more from the cognitive load incurred by the presence of competing speech signals. Furthermore, non-native listeners might not be able to rely on other sources of information (e.g., lexical cues) to compensate for the depletion of processing resources caused by concurrent tasks (Mattys, Carroll, Li, & Chan, 2010), because of their deficits in exploiting high-level linguistic cues.

While the disproportionate effect of cognitive load or informational masking on L2 speech recognition can arise due to L2 listeners' inability to exploit language-specific cues, the effect can be greater for non-native listeners also because listening to L2 speech requires greater processing resources or listening effort[18] by itself, thereby depleting their processing resources. Based on the ELU model (e.g., Rönnberg, 2003), this is because the degree of a mismatch between an incoming acoustic signal and the listener's mental representations is expected to be larger for non-native listeners, whose L2 phonological representations are less precise and may deviate from the acoustic input (e.g., their representations are influenced by their L1 knowledge). Furthermore, non-native listeners' linguistic processes are less developed for the target language, thereby requiring greater processing load compared to L1 processing which involves more automatized processes (see Clahsen & Felser, 2006 for a review).

[18] The term 'listening effort' can refer to perceived (i.e., subjective) effort, but more objectively, it means the amount of processing resources allocated to a task, that is, processing load (Lemke & Besser, 2016).

Neuroimaging studies have shown greater activations for L2 than L1 speech processing in some language areas such as the left inferior frontal gyrus, or distinct activations for L2 processing in areas that are not typically used for L1 processing, such as the anterior cingulate (e.g., Callan, Jones, Callan, & Akahane-Yamada, 2004; Dehaene et al., 1997; see Indefrey, 2006; Stowe & Sabourin, 2005 for reviews). Similarly, previous electrophysiological studies have shown that L2 listeners are slower with certain linguistic processes (e.g., semantic or syntactic integration processes) than are native listeners (e.g., Hahne, 2001; Hahne & Friederici, 2001; see Newman, Tremblay, Nichols, Neville, & Ullman, 2012; Weber-Fox & Neville, 1996 for N400 during reading), and that they may not be able to attain a native-like automatic process (e.g., left anterior negativity in response to morphosyntactic violations) even if they are highly proficient in their L2 (e.g., Mueller, Hahne, Fujii, & Friederici, 2005; Mueller, 2005). A recent pupillometry study by Schmidtke (2014) examined cognitive effort needed for lexical retrieval during spoken word recognition; bilingual listeners had an overall delayed pupil response compared to monolingual listeners, and a neighbourhood density effect (i.e., greater retrieval effort for words with higher neighbourhood density) was greater for bilinguals than for monolinguals. Effects of word frequency and neighbourhood density also varied among bilingual listeners depending on their language proficiency (i.e., smaller effects with increasing proficiency). Taken together, these findings indicate that due to their intrinsic L2 speech recognition problems that require additional processing resources, L2 listeners have less cognitive resources left to deal with cognitive demands of adverse listening conditions.

4.1.4 Neural measures of auditory and lexical processing

The present study investigated speech processing by L1 and L2 listeners in two-talker situations where listeners must selectively attend to the desired talker while ignoring the distracting talker. As discussed above, the presence of concurrent speech streams increases the cognitive demands of the listening situation. The neural measures described in the current section (4.1.4) provided a means to examine auditory and lexical processing by L1 and L2 listeners in this listening situation. Specifically, the present study examined cortical entrainment to the speech envelope and the N400 response.

4.1.4.1 Cortical entrainment to the amplitude envelope of speech

While multi-talker environments make speech recognition harder, humans as well as other animal species possess the ability to attend to one auditory object with relative ease by decomposing the complex auditory scene into separate auditory objects (e.g., Bregman, 1990). Researchers have sought to find the neural mechanisms underlying this “cocktail-party effect” (Cherry, 1953). Recent studies have shown that when listeners selectively attend to a target talker against distracting talkers, low-frequency entrainment to the speech envelope is relatively enhanced for that target speech (Ding & Simon, 2012a; Kerlin et al., 2010; Zion Golumbic et al., 2013; Horton, D’Zmura, & Srinivasan, 2013), revealing neural mechanisms by which the auditory system selectively processes target speech streams in a complex auditory environment.

Specifically, Ding and Simon (2012a) reconstructed the temporal envelopes of target and distracting talkers using MEG signals that were recorded while listeners attended

to one of the two competing talkers. The reconstructed envelope was more strongly correlated with the envelope of the target talker than that of the distracting talker or the two talkers combined. This suggests that entrainment to the amplitude envelope of speech can be modulated by top-down attention in a complex auditory scene. Moreover, neural responses to the attended speech were found to adapt to the intensity of that signal, suggesting object-based intensity gain control. Zion Golumbic et al. (2013) also examined neural tracking in competing-talker environments, by measuring low-frequency phase and high gamma power in direct recordings made from the surface of the cortex (i.e., ECoG); robust entrainment was found for both target and distracting signals in low-level auditory regions (i.e., superior temporal gyrus) although the response to the attended speech was relatively enhanced. In contrast, "selective" entrainment to the target talkers was also observed (i.e., no detectable response to the unattended speech) in higher-order language and attentional control regions as well as low-level auditory areas.

Furthermore, previous work has found a clear interaction between speech intelligibility and attention in envelope-tracking activity (for further discussion of the relationship between speech intelligibility and entrainment, see Chapter 3). For example, Rimmele et al. (2015) found more robust entrainment for attended than ignored sentences, but only when the sentences were natural (i.e., no differences between attended and unattended sentences when they were noise-vocoded). The authors argued that the enhanced entrainment observed for more intelligible, attended speech indicates that listeners exploited higher-level linguistic information to aid lower-level auditory tracking of the speech envelope.

Similarly, Kong et al. (2015) found that the difference in neural entrainment between attended and ignored speech streams increased as the spectral resolution of the speech signals increased (i.e., unprocessed speech and noise-vocoded speech generated with varying numbers of channels). They also found a significant correlation between listeners' attentional modulation of the response (i.e., the entrainment difference between attended and unattended speech signals) and speech comprehension performance. In contrast to the explanation provided by Rimmele et al. (2015), the authors of this study suggested that this likely occurred because speech segregation cues were less available due to spectral degradation, thereby impairing listeners' ability to employ top-down attention to selectively process the target speech over the distractor. Although the exact causal relationship between attentional control of neural speech tracking and speech comprehension remains unknown, it seems that top-down attentional modulation of selective tracking is correlated with speech intelligibility.

Furthermore, a recent EEG study by O'Sullivan et al. (2015) found that target-talker entrainment in a two-talker situation was correlated with performance on a high-level attention task. Specifically, there was a significant positive correlation between neural selectivity for attended speakers (i.e., the accuracy of determining attended speakers using the stimulus reconstruction method, similar to as used in Ding and Simon, 2012a) and how accurately subjects answered questions about the stories that they heard. Because the subjects were young normal-hearing listeners (i.e., this study did not test non-native listeners in particular), their performance on the behavioural task likely depended on how well they attended to the target talkers and remembered what they heard, rather than their linguistic knowledge or language experience. The

correlation seen in this study thus demonstrates that neural tracking of the speech envelope in a complex auditory scene is directly modulated by listeners' attentional deployment during the task. Focusing more attention on the target signals may have also facilitated speech comprehension in this difficult listening condition. However, exerting more attentional effort may not necessarily lead to better speech comprehension especially for non-native listeners, who have relatively poor L2 knowledge. That is, greater focused attention on the target speech may reflect greater comprehension difficulties or listening effort experienced by listeners.

In the present comparison of L1 and L2 listeners, one could expect that L1 listeners would have greater target-talker entrainment than L2 listeners, based on the links between speech comprehension and selective neural tracking of attended speech (Rimmele et al., 2015; Kong et al., 2015). Specifically, L1 listeners might have greater target-speech entrainment because it has been thought to be enhanced by top-down prediction using linguistic cues in some studies (Rimmele et al., 2015; also see Peelle and Davis, 2012 for a review). It is equally plausible that native listeners are better at segregating competing speech streams than are non-native listeners (e.g., Cooke et al., 2008) and thus build a more robust neural representation of the target speech signal over the distractor (Kong et al., 2015). In contrast, it is possible that target-talker entrainment is modulated by listening effort or attention needed for the task (e.g., O'Sullivan et al., 2015), independently of listeners' speech comprehension. If the latter is the case, L2 listeners might have greater target-talker entrainment because they likely need greater cognitive resources to understand target speakers than do L1 listeners.

4.1.4.2 N400

The N400 component of the event-related brain potential (ERP) is a negative response that peaks at around 400 ms after word onset, which was first known as a negative response to semantic incongruity (e.g., "*He spread the warm bread with socks*"; Kutas & Hillyard, 1980). N400 is thought to be a marker of lexical and semantic processing which reflects the amount of effort spent integrating the target word into previous context (e.g., Brown & Hagoort, 1993; Hagoort, 2008; Osterhout & Holcomb, 1992); words that are predictable from the context require less processing effort (thus reduced N400) than those that are harder to predict. The N400 response normally has a centro-parietal maximum (see Kutas & Federmeier, 2011 for a review) and can also reflect the ease of lexical access (Federmeier, 2007; Kutas & Federmeier, 2000). According to this view, the N400 response does not necessarily mirror a combinatorial process in which a lexical item is integrated into the preceding context. Instead, any factors that facilitate lexical access such as word frequency can lead to a reduction in N400 amplitude (e.g., Van Petten & Kutas, 1990; for a review, Lau, Phillips, & Poeppel, 2008). Predictable words are thus easier to access from long term memory (i.e., smaller N400) because the context pre-activates features related to that lexical item.

Researchers have examined the N400 response to investigate how listeners modulate their lexical and semantic processing in adverse listening conditions. Aydelott, Dick, and Mills (2006) found that low-pass filtering the sentence context reduced N400 differences between congruent and incongruent final words, which was driven by decreased N400 amplitudes for incongruent words. Similarly, Obleser and Kotz (2011) found that N400 differences between high and low cloze probability keywords became

smaller with increasing levels of signal degradation (i.e., noise-vocoded sentences). Specifically, N400 amplitudes for low cloze keywords monotonically increased with better signal quality, whereas N400 for high cloze keywords showed an inverted-u shape pattern (i.e., greater N400 for the 4-channel than 1-channel or 16-channel conditions). In addition, the N400 effect peaked earlier for more intelligible conditions. These findings suggest that lexical semantic processing is disrupted when the signal quality does not allow listeners to use contextual cues in the speech signal (cf. Boulenger, Hoen, Jacquier, & Meunier, 2011; Strauß, Kotz & Obleser, 2013). Informational masking can also affect the N400 response; Carey, Mercure, Pizzioli, and Aydelott (2014) found that the overall magnitude of the N400 response was reduced regardless of semantic conditions when listeners attended to one of two competing speakers that were presented dichotically. The authors suggested that this occurred because informational masking disrupted the engagement of speech comprehension processes.

The N400 response can be useful for the purpose of the present work, also because N400 has been explored for a range of native and non-native talkers and listeners in previous research. Accented speech can also be seen as one form of degraded speech, in that it contains phonetic and phonological features that deviate from listeners' mental representations. Listeners thus need a greater amount of cognitive effort to process accented speech (e.g., Van Engen & Peelle, 2014). In addition, listeners need a normalisation or adaptation process to be able to correctly map segmental and suprasegmental deviations in foreign-accented speech onto their phonetic/phonological representations (e.g., Bradlow & Bent, 2008; Clarke & Garrett,

2004). Behavioural studies have consistently found that listeners are less accurate and slower at comprehending accents that are different from their own (e.g., Bent & Bradlow, 2003; Adank et al., 2009; see Chapter 2 for further discussion), but it remains unclear how the phonetic/phonological variability caused by speaker accent affects semantic integration or lexical access as measured by N400.

In Goslin, Duffy, and Floccia (2012), N400 amplitude in response to low cloze probability words was smaller for foreign accents than for unfamiliar regional accents or the listeners' own native accent. However, Hanulíková, van Alphen, van Goch, and Weber (2012) reported similar N400 effects (i.e., N400 difference between semantically correct and incorrect sentences) in response to a native Dutch accent and Turkish-accented Dutch, although the response was more widely distributed in the scalp for the foreign-accented speech (i.e., anterior and posterior distribution). In contrast, Romero-Rivas et al. (2015) found that semantic violations elicited larger N400 amplitudes in foreign- than native-accented speech. Foreign-accented speech also elicited a more negative N400 response for words in semantically correct sentences than did native-accented speech, but this difference disappeared in the second block. The authors suggested that this was likely because listeners learned to use lexical-semantic information to adapt to foreign-accented speech (i.e., a reduction in N400 amplitude was only found for semantically correct words). They also found more widely distributed N400 responses for foreign-accented speech compared to the more-typical centro-parietal distribution of the response for native speech, similar to Hanulíková et al. (2012). The authors argued that this topographic difference might

suggest that listeners needed greater cognitive resources to process foreign-accented speech.

These inconsistencies among previous studies could have occurred because lexical and semantic processing can have a complex relationship with speech intelligibility; listeners can increase their reliance on contextual cues or exert additional lexical processing effort when the speech signal becomes less intelligible (e.g., accented speech or acoustically degraded speech; e.g., Kalikow et al., 1977; Miller et al., 1951; Obleser, Wise, Dresner, & Scott, 2007). However, lexical processing can be hindered, if the acoustic signal is severely degraded and can thus not be sufficiently decoded (Obleser & Kotz, 2010, 2011; Obleser, Wise, Dresner, & Scott, 2007).

Previous studies have also examined N400 in L2 listeners; Hahne and Friederici (2001) found similar N400 effects in native listeners and late L2 learners (i.e., their average age of learning was 21), although the response of the L2 listeners was relatively delayed. In Hahne (2001), however, L2 listeners who had started learning the L2 after the age of 10 had a N400 response that was larger and delayed for semantically correct sentences compared to that of L1 listeners (i.e., thus smaller N400 effects in L2 than L1 listeners) as well as being extended to frontal electrodes. No difference was found for semantically incorrect sentences. These results indicate that L2 listeners may not be unable to attain native-like semantic processing when they reach a certain-level of L2 proficiency, but that they likely need greater effort for lexical processing even when words are easily predictable from context.

In the present study, L2 listeners' lexical-semantic processing can be relatively disrupted compared to L1 listeners, due to their deficits in making use of semantic-contextual cues or failure to map the acoustic input into correct lexical representations. That is, facilitation and inhibition of words using contextual information may be hindered during L2 speech processing, resulting in diminished N400 effects (i.e., smaller N400 differences between predictable and less predictable conditions; Hahn, 2001). This is also expected because L2 speech recognition is cognitively more demanding; similarly, N400 effects were shown to be attenuated under higher working memory load (D'Arcy, Service, Connolly, & Hawco, 2005; Gunter, Jackson, & Mulder, 1995). It is also possible that L2 listeners might exhibit greater (i.e., more negative) N400 amplitudes compared to L1 listeners, if they exert additional effort for lexical processing to compensate for their L2 speech recognition difficulties.

4.1.5 Aims of the present study

The aim of the current study was to investigate how speech recognition difficulties experienced by non-native listeners in two-talker situations affect their auditory and lexical processing. To this end, this study compared neural tracking (i.e., entrainment to the speech envelope) and N400 responses for target talkers that were recorded in L2 listeners to those recorded in L1 listeners in an EEG experiment. The subjects were also asked to perform a behavioural task to assess their speech comprehension accuracy. In this study, target and distracting speakers were presented to separate ears. Therefore, information masking was expected to occur without energetic masking as listeners could hear the target speech signal clearly through one ear once segregated from the distracting signal. As discussed above (Chapter 4.1.4), L2 listeners' auditory

and lexical processing can be more severely impaired under this adverse condition (e.g., cognitive load) than that of L1 listeners, but it is possible that additional listening effort that arises due to their inadequate linguistic knowledge combined with the demands of the listening condition results in enhancement in these processes as a compensatory mechanism.

In addition, the current study varied the accent of the speech materials to determine how these processes are modulated by whether or not the accent of the listeners matches that of the talkers, which is another real-life factor which affects speech intelligibility (see Chapter 2 for details). Listeners likely need greater lexical processing when listening to accents that do not match their own accent (e.g., Romero-Rivas et al., 2015), but results of this kind are not always found (e.g., Goslin et al., 2012). It is also possible that less intelligible accents attenuate the size of N400 effects by hindering the use of contextual cues. Accent can also affect neural entrainment to the target speakers; target speakers with more intelligible accents might facilitate selective entrainment to the target speech (e.g., Rimmele et al., 2015), or conversely, difficult accents may cause listeners to focus more attention on the target signal, thereby enhancing the target-talker enhancement.

In this study, native English and Korean listeners (i.e., L2 listeners) heard pairs of simultaneous English sentences spoken in two different accents (Standard Southern British English and Korean) and presented to separate ears. EEG was recorded while listeners were instructed to selectively attend to one of the talkers. Neural entrainment was measured as the amount of phase coherence between EEG signals and the

amplitude envelope of the speech from the target and distractor talkers. This study used sentences that differed in terms of the predictability of the final word, which allowed for lexical processing to be simultaneously assessed (i.e., N400). Subjects were instructed to press a button whenever they heard a semantically anomalous sentence in the target ear (i.e., catch trials), and the accuracy of the button response was used as a behavioural measure of their speech recognition performance.

4.2 Methods

4.2.1 Subjects

Twenty-three native speakers of British English (12 female) and 21 native speakers of Korean (14 female) participated in the experiment. The British subjects reported that they were native speakers of Standard Southern British English, except for 3 subjects who grew up in other parts of England (South West or Northern England). All the British subjects were monolingual speakers. The Korean subjects reported that they had started learning English at school in South Korea at an average age of 10 years old (5-14 y), and that they had not lived in English-speaking countries before they became adults. Their average length of residence in English-speaking countries as adults was 1 year (1-31 months). One British subject and two Korean subjects were excluded from the analyses because they had noisy recordings; they had several bad channels or less than 50% of trials left after artefact rejection. All subjects were right-handed adults under 35 years old (mean: English = 21.8 y, Korean = 26.5 y) without any self-reported hearing or neurological impairments.

4.2.2 Stimuli

English sentences were recorded by female native speakers of Standard Southern British English and Korean (i.e., one speaker each; they were both 28 years old). The Korean speaker reported that she had been living in London for one year at the time of testing. The stimuli consisted of 720 pairs, each consisting of two sentences produced by each of the talkers. Sentences within a pair were matched in duration. The average duration of the British sentences was originally 0.44 seconds shorter than that of the Korean speaker, so the sentences of the British speaker were lengthened and those of the Korean speaker were shortened by 10% using a pitch-synchronous-overlap-and-add (PSOLA) procedure (Boersma & Weenink, 2014). All the stimuli had 44100 16-bit samples per second. The stimuli were counterbalanced between subjects and the order of presentation was randomized. Sentences presented in the target ear were not repeated.

The sentences varied in the predictability of the final word to allow for measurement of N400. This study used an existing corpus of N400 stimuli designed for L2 learners (Stringer, 2015), and expanded the number of sentences by editing another L2 sentence corpus (Calandruccio & Smiljanic, 2012) to vary final-word predictability (Appendix 1). High cloze probability sentences comprised 42.5% of the stimuli. They were made up of strongly constraining sentence contexts and congruent final words as in *Beef and milk come from cows*. Another 42.5% of the stimuli were low cloze probability sentences, neutral sentences such as *The man draws pictures of cows*. The remaining 15% of the stimuli was made up of semantically anomalous sentences, which had

strongly constraining sentence contexts but ended with incongruent words, as in *Beef and milk come from bays*.

Table 4-1: Example sentences from Stringer (2015) that were used in the experiment. Sentences differed in the predictability of the final word.

Semantic condition	Sentence examples
High cloze probability sentences	*Patients are cared for by doctors and nurses.*
Low cloze probability sentences	*Trains and buses have big wheels.*
Anamolous sentences	*Wine is usually made from wool.*

4.2.3 Apparatus

All stimuli were presented via Praat (Boersma & Weenink, 2014) using an external sound card (RME Fireface UC) which was connected to a custom-built trigger box. The trigger box was used to deliver the stimuli of each speaker separately to left and right channels via Etymotic ER-1 insert earphones. To obtain timing information of the stimuli, triggers were generated as pulses on a separate audio channel, which were converted to TTL triggers via a custom circuit.

EEG was recorded through a Biosemi Active Two system with 64 (Ag/AgCl) electrodes mounted on an elastic cap and 7 external electrodes (left and right mastoids, nose, two vertical and two horizontal EOG electrodes). Unreferenced EEG signals were recorded with a sampling rate of 2048 Hz. Electrode impedances were kept within the range of ±25k Ω during the experiment. Time-aligned triggers were also recorded by the EEG system.

4.2.4 Procedure

During EEG recording, the sentences were presented simultaneously in different ears with a different talker in each ear. Subjects selectively attended to a target ear/talker and pressed a button whenever they heard a semantically anomalous sentence in that ear. Before each block started, subjects were told which talker they were to attend to via which ear. In addition, an Asian or white Caucasian female face lit up in the left or right side of a small tablet placed in front of the subjects to ensure that they attended to the correct talker. Subjects were given a short break between blocks. The target talker and the ear in which her speech was presented alternated every block. The experiment consisted of 8 blocks of 90 stimuli (i.e., 90 sentence pairs). The duration of inter-stimulus silence intervals was randomly jittered from 1.5 to 1.7 seconds.

4.2.5 Analysis

4.2.5.1 Pre-processing

After recording, the EEG signals were referenced to the average of the left and right mastoids. Noisy channels were interpolated. The data were then high-pass filtered at 0.1 Hz and low-pass filtered at 40 Hz using Butterworth filters as implemented in the ERPlab toolbox (Lopez-Calderon & Luck, 2014) of EEGlab (Delorme & Makeig, 2004). Independent Component Analysis was also applied to the data to remove components containing eye blinks and horizontal eye movements. All pre-processing procedures, except for filtering, were performed in Matlab using the Fieldtrip toolbox (Oostenveld et al., 2011).

4.2.5.2 N400 analysis

To measure the N400 response, the data were segmented into epochs time-locked to the onset of each final word (200 ms pre-stimulus and 1000 ms post-stimulus intervals). Trials with amplitude exceeding ±150 μV were rejected, and the rejection rate averaged across subjects was 12.6 %. After subtracting the baseline average over the pre-stimulus interval, N400 amplitudes were measured by averaging the amplitude in the 300-500 ms time window. N400 amplitudes were averaged across five midline electrodes, Fz, FCz, Cz, CPz and Pz, similar to the previous N400 literature (e.g., Strauß et al., 2013).

4.2.5.3 Coherence analysis

The degree of phase-locking was measured between EEG signals and the amplitude envelope of speech for both target and distracting talkers using 'coherence' (i.e., cerebro-acoustic coherence, Peelle et al., 2013). Before computing coherence, amplitude envelopes were calculated from the speech stimuli using full-wave rectification and filtering (i.e., high-pass filtered at 0.1 Hz and low-pass filtered at 40 Hz using Butterworth filters from ERPlab). The amplitude envelopes were also down-sampled to 2048 Hz to match the sampling rate of the EEG data. The continuous EEG signals and amplitude envelopes were segmented into 2-second Hanning-windowed epochs that were time-locked to the onset of each sentence. Coherence between the amplitude envelope and EEG signals was calculated from the cross-spectral density of the FFT of the two signals, divided by the power spectrum of each signal (see Chapter 3.2.5.2 for details).

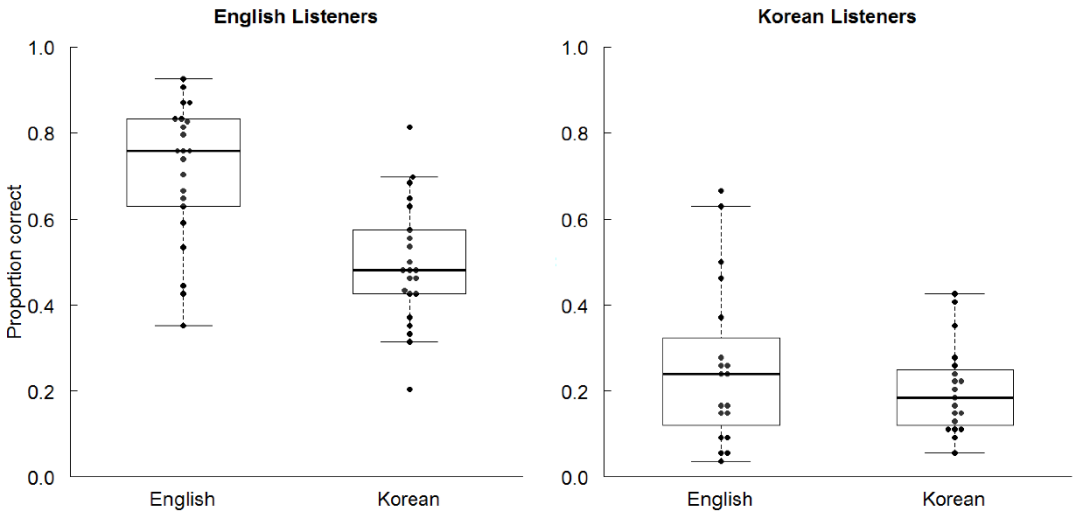
Denoising Source Separation (DSS; de Cheveigné & Simon, 2008) was used to isolate the neural activity that was phase-locked to the amplitude envelopes of the stimuli. Specifically, DSS components were extracted based on spatial filters (i.e., linear combinations of the electrodes) that were calculated from the covariance of the raw data at each electrode and the covariance of the coherence calculation at each electrode over all trials (see Chapter 3.2.5.3 for details of this technique). The DSS components were calculated across conditions, rather than specifically extracting activity that had different entrainment for targets and distractors. Related applications of DSS have been shown to be effective in isolating the envelope-tracking response (e.g., Ding & Simon 2012a; Ding et al., 2016). The current study used the first four DSS components that maximized the reliability of coherence for each subject; it appeared that the first four components all captured activity related to envelope tracking, but with varying degrees. The components were then projected back into sensor space.

4.3 Results

The topographies in Figure 4-2 display the coherence values for target talkers in the delta-theta range (1-8 Hz) averaged across conditions of speaker accent and ear of presentation for each listener group. The tracking activity was mostly found in frontocentral electrode sites, which is in agreement with the previous finding that this response originates from bilateral auditory cortex (e.g., Luo & Poeppel, 2007; Doelling et al., 2014). However, a direct comparison with other studies such as MEG or fMRI studies is difficult to make without performing source localisation which would be much more difficult with EEG data than with fMRI or MEG data (Luck, 2005). For statistical analysis, coherence values were averaged across frontocentral electrodes

(F1, F2, F3, F4, F5, F6, F7, F8, FCz, FC1, FC2, FC3, FC4, FC5, FC6, FT7, FT8, Cz, C1, C2, C3, C4, C5, C6). In this study, mixed-model analyses were carried out without performing model selection; all relevant fixed factors and their interactions were included in the models as well as random intercepts, based on previous findings and the aims of the current study[19].

Figure 4-1: Combined boxplot and beeswarm plots of the proportion of correctly identified anomalous sentences by speaker accent (English and Korean) for English (L1) and Korean (L2) listeners.



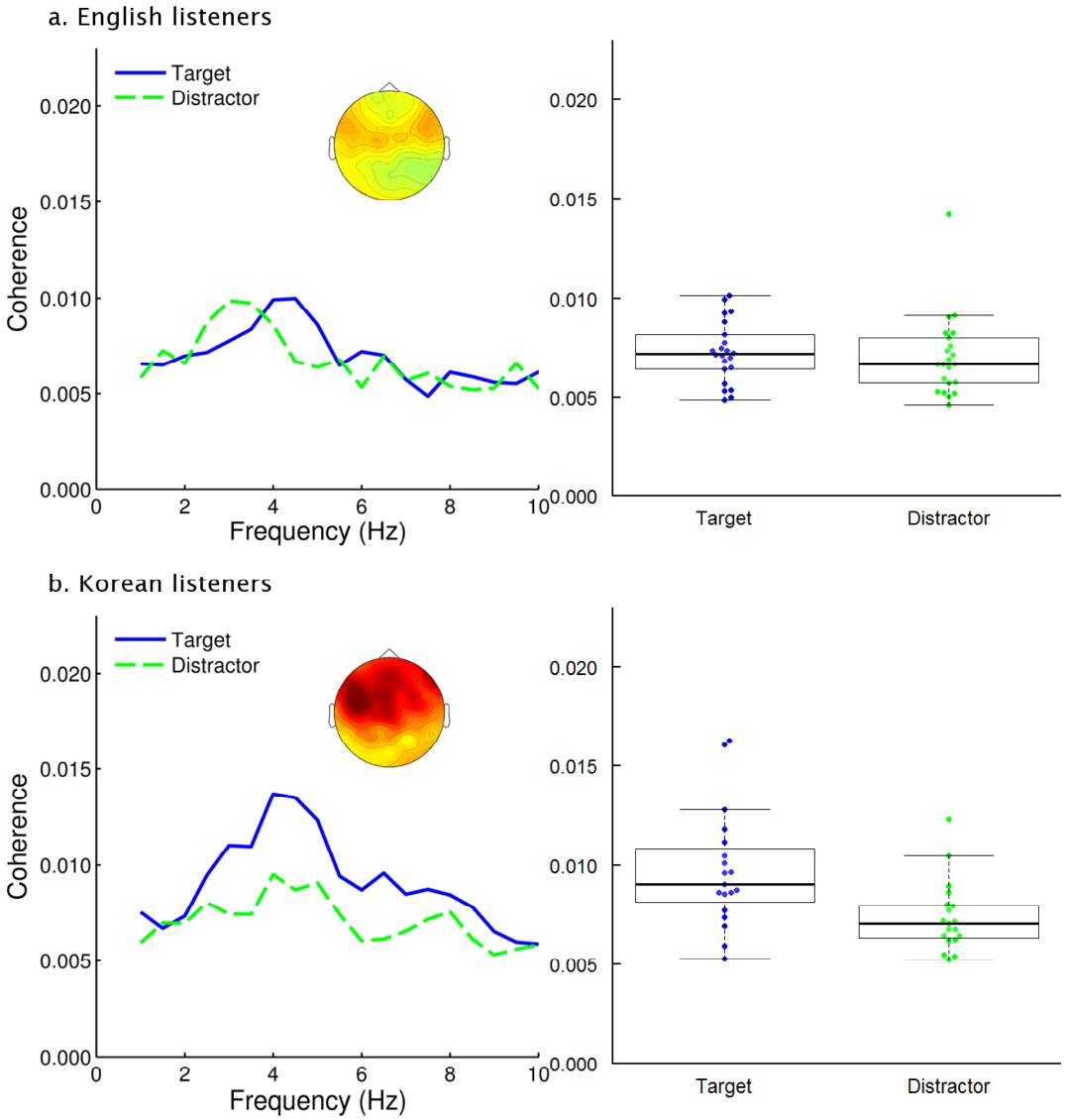
As displayed in Figure 4-1, English listeners gave more accurate behavioral responses (i.e., identification of anomalous sentences) than did Korean listeners. Moreover, English listeners had an intelligibility advantage for English-accented speech compared to Korean-accented speech, whereas Korean listeners' performance was fairly similar for both accents. False alarm rates (i.e., button presses to non-anomalous sentences) were very low (mean; English listeners = 2 %, Korean listeners = 3 %) and

[19] The results were highly similar when the model selection approach was used to find the best-fitting models.

were thus not analyzed. Statistical analyses were performed using the R package lme4 (Bates et al., 2015). The package CAR (Fox & Weisberg, 2002) was used to calculate type II analysis-of-variance tables. A logistic mixed-model analysis was performed on behavioral responses, using the button response to anomalous sentences (i.e., correct or incorrect) as the dependent variable; listener group (i.e., English and Korean listeners), ear of presentation (i.e., left and right), and speaker accent (i.e., Standard Southern British English and Korean-accented English) as independent variables; and random intercepts for each subject and sentence stimulus. The results verified that there were main effects of listener group, $\chi^2(1) = 70.94$, $p < 0.001$, and speaker accent, $\chi^2(1) = 27.78$, $p < 0.001$, and a significant interaction between these two variables, $\chi^2(1) = 20.98$, $p < 0.001$. Although a right-ear advantage can be expected for speech processing because of the dominant contralateral pathway from the right ear to the left hemisphere (e.g., Kimura, 1961), there was no significant effect of ear, $p = 0.0779$ in this case, similar to Carey et al. (2014).

Despite the fact that Koreans found this task harder, the coherence results demonstrate that Korean listeners actually had greater entrainment to target talkers than did English listeners (Figure 4-2). That is, both listener groups had coherence peaks in the delta-theta range (1-8 Hz) across conditions of speaker accent and ear of presentation, but only L2 listeners had greater coherence for the target talker. In addition, the topographies displayed in Figure 4-2 suggest that this enhanced coherence for target talkers by L2 listeners was left-lateralized, although source localisation was not performed in the current study.

Figure 4-2: Results of the coherence analysis for English (L1) and Korean (L2) listeners. Coherence values averaged across conditions of speaker accent and ear are plotted as a function of frequency (0-10 Hz) with topographies of the mean coherence values for target talkers in the delta-theta range (1-8 Hz; left). Combined boxplot and beeswarm plots of individual coherence values (right).



A mixed-model analysis was conducted with coherence values averaged in the relevant frequency range (i.e., 1-8 Hz) as the dependent variable; listener group, target type (i.e., target and distractor), ear of presentation, and speaker accent as independent variables; and with by-subject random intercepts. The interaction between listener group and target was significant, $\chi^2(1) = 14.77$, $p < 0.001$, as well as the main effects of listener group, $\chi^2(1) = 5.45$, $p = 0.020$, and target, $\chi^2(1) = 11.61$, $p < 0.001$. However, there was no significant effect of ear, $p = 0.751$, showing that entrainment for the right-ear presentation did not differ from that for the left ear. The current study focused on the coherence results for the whole delta-theta range (1-8 Hz), but the mixed-model analysis for the theta range (4-8 Hz) reproduced the same significant effects.

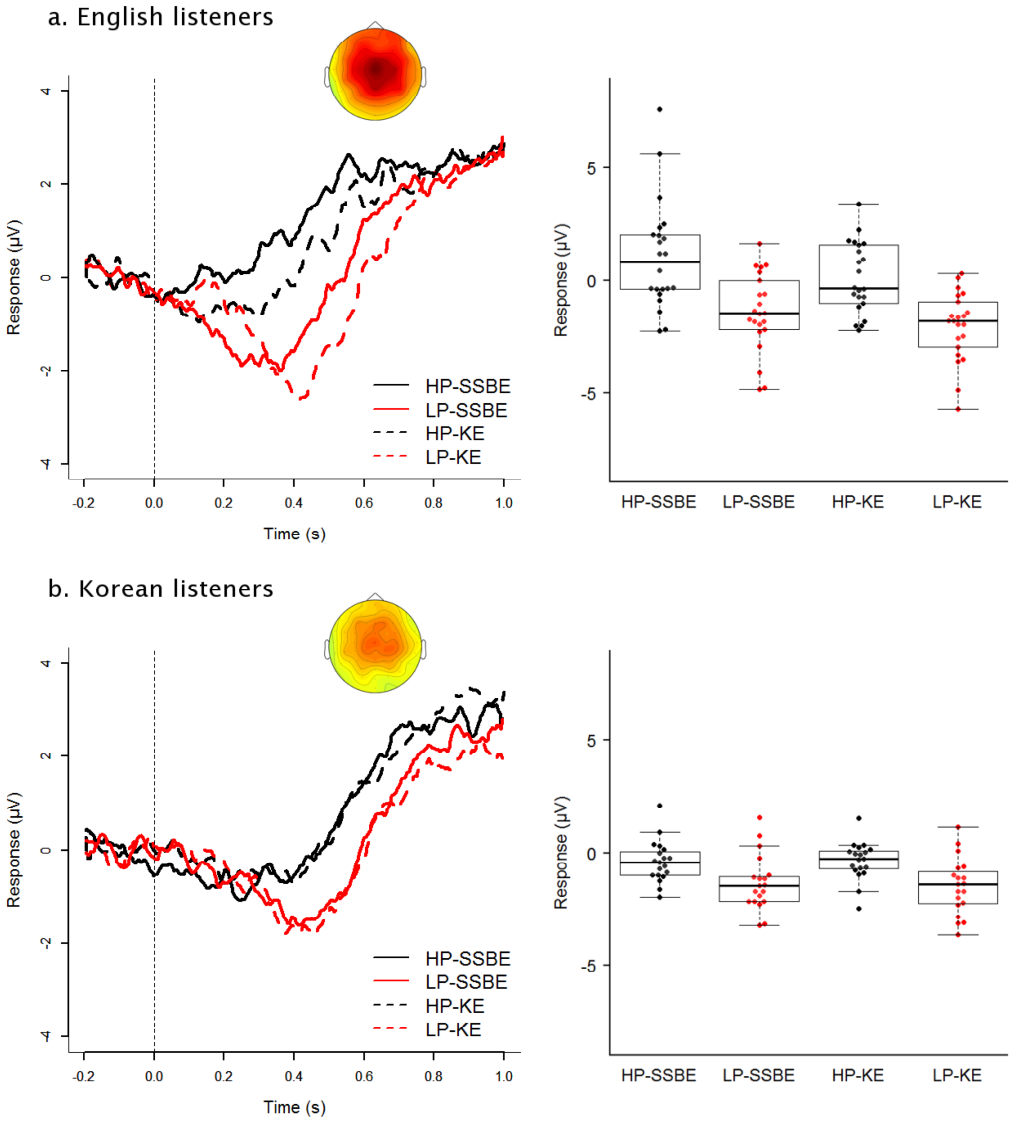
It thus appears that L2 listeners responded to their greater recognition difficulty with greater focused attention, which enhanced their neural tracking of the target speech signal. This result is the opposite of previously reported positive relationships between envelope tracking and intelligibility (e.g., Peelle et al., 2013 for single-talker environments; Rimmele et al., 2015 for competing-talker environments). In contrast, L1 listeners had no significant increase in target-talker entrainment, likely because the listening condition was easier than in previous studies. That is, competing speakers were presented in different ears in the current study rather than mixed together as in Ding and Simon (2012a). It thus appears that they were able to perform the task with little additional focused attention.

As discussed above, English listeners were more accurate at detecting anomalous sentences in the English accent than the Korean accent, while Koreans had smaller

differences between the accents. However, these behavioural differences in accent processing were not observed in the neural entrainment results; there were no significant main effects or interaction for speaker accent, $p > 0.05$. This may have occurred because the Korean accent was still not difficult enough for L1 listeners to require greater focused attention (i.e., they still understood Korean-accented speech better than did Korean listeners), and Koreans found both accents similarly hard.

A mixed-model analysis was performed for the N400 for the final word in each target-talker sentence, with N400 amplitudes included as the dependent variable; listener group, speaker accent, sentence type (i.e., high cloze and low cloze probability sentences), and ear of presentation as independent variables, and with by-subject random intercepts. As displayed in Figure 4-3, the results demonstrated a typical N400 effect, with greater amplitudes in low than high cloze probability sentences, suggesting that listeners exerted more effort for processing the final word when it was less predictable. The main effect of sentence type was significant, $\chi^2(1) = 100.43$, $p < 0.001$. Carey et al. (2014) found greater N400 effects when sentences were presented to the left ear; the authors claimed that this indicates that the right hemisphere is more involved with "integrating" the target word into the preceding context, whereas the left hemisphere is more involved with "predicting" upcoming words based on the context (Wlotko & Federmeier, 2007). However, ear was not significant in the current study, $p = 0.136$.

Figure 4-3: Results of the N400 analysis for English (L1) and Korean (L2) listeners. Grand average ERPs (N400) for sentence-final words by sentence type (HP: high cloze probability sentences, LP: low cloze probability sentences) and speaker accent (SSBE: Standard Southern British English, KE: Korean-accented English) are plotted for English (L1) and Korean (L2) listeners, with topographies of the mean N400 differences between HP and LP sentences (left). Combined boxplot and beeswarm plots of individual N400 values (right).

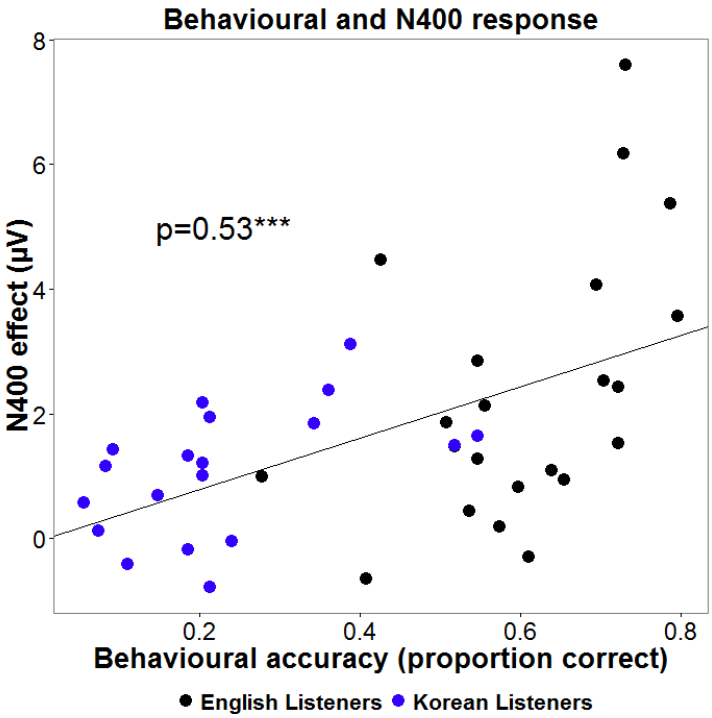


English listeners had significantly greater differences between low- (LC) and high-cloze (HC) sentences than did Korean listeners; the interaction between sentence type and listener group was significant, $\chi^2(1) = 12.23$, $p < 0.001$. Specifically, this difference between English and Korean listeners emerged largely from the high cloze

probability condition (i.e., mean amplitudes; English Listeners * HC = 0.57; Korean Listeners * HC = - 0.34; English Listeners * LC = - 1.73; Korean Listeners * LC = - 1.43). The results also revealed that the N400 was modulated by the accents of the talkers and listeners. Specifically, Korean listeners had similar N400 amplitudes for both accents, whereas English listeners had larger N400 amplitudes for Korean-accented English than for Southern British English regardless of whether or not the final word was highly predictable. The two-way interaction between listener group and speaker accent, $\chi^2(1) = 3.83$, $p = 0.050$, and the main effect of speaker accent were significant, $\chi^2(1) = 6.45$, $p = 0.011$. This suggests that lexical processing as reflected by N400 more closely mirrored the behaviorally measured intelligibility of these sentences. That is, English listeners needed additional lexical processing to compensate for the less-intelligible Korean accent, and Korean listeners needed effortful processing for both accents particularly in the highly predictable condition.

To further investigate the relationships between these responses, correlation analyses were carried out for individual listeners' average button response accuracy, target-talker selectivity (i.e., difference in coherence between target and distracting talkers), and N400 effect (i.e., difference in N400 between HP and LP conditions). As displayed in Figure 4-4, there was a significant correlation between the N400 effect and button response accuracy for the entire group of subjects, $r = 0.53$, $p < 0.001$. The correlation was also significant within subject groups (i.e., English listeners: $r = 0.50$, $p = 0.019$; Korean listeners: $r = 0.49$, $p = 0.032$). That is, individuals who were more accurate at detecting anomalous sentences tended to have a larger N400 effect, suggesting a direct link between speech intelligibility and N400.

Figure 4-4: Correlation between listeners' behavioural performance (i.e., button presses to anomalous sentences in the target ear) and N400 effect. The solid line represents a regression line.



In contrast, target-talker selectivity was only significantly correlated with behavioral accuracy when calculated across listener groups, r = -0.35, p = 0.026, not within English, r = -0.13, p = 0.560, or Korean groups, r = -0.03, p = 0.895. This appears to reflect group-level differences between Korean and English listeners, with Korean listeners having larger target-talker selectivity, but lower behavioural accuracy compared to English listeners. Likewise, target-talker selectivity was only significantly correlated with N400 effect for the entire group of subjects, r = -0.36, p = 0.020, but not within English, r = -0.36, p = 0.100, or Korean groups, r = -0.07, p = 0.762. This also seems to mirror group-level differences with Korean listeners having larger target-talker selectivity and smaller N400 effects compared to English listeners. The results thus suggest that target-talker entrainment is not a simple function of individuals' recognition difficulties or lexical processing effort.

4.4 Discussion

It has long been obvious that L2 speech comprehension is more effortful, and this is particularly so in adverse listening conditions such as multi-talker environments. Surprisingly, the results demonstrated that L2 listeners had more selective neural entrainment to target talkers than did L1 listeners, likely because they needed greater listening effort during the task. Furthermore, L2 listeners had additional lexical processing (i.e., larger N400 amplitudes) for words in highly predicable contexts than did L1 listeners, causing a relative convergence of N400 amplitudes between high and low cloze predictability sentence conditions. In contrast, L1 listeners responded to the speech recognition difficulties caused by an L2 accent only at the lexical level by increasing the overall degree of lexical processing (i.e., no selective neural entrainment).

It was expected that native listeners would have more robust target-speaker entrainment than L2 listeners, given previous work finding greater selective entrainment for more intelligible speech (Rimmele et al., 2015; Kong et al., 2015). However, the present work demonstrated that cortical entrainment to speech can also be greater when intelligibility is lower, even when the speech is heard by L2 listeners who have less developed higher-level linguistic processes for that language. It appears that this occurred because L2 listeners found the speech recognition task more difficult, and thus deployed a greater amount of cognitive effort or attention.

This supports the previously found positive relationship between target-talker entrainment in two-talker situations and performance on a high-level attention task

(O'Sullivan et al., 2015). The present results, as well as those of O'Sullivan et al. (2015), are interesting, because they indicate that an increase in neural entrainment to speech in competing-talker environments can reflect listening effort, which can vary depending on the difficulty of the task. Furthermore, it is possible that this additional auditory processing seen in the current study played a compensatory role in speech recognition by facilitating speech segregation or later speech comprehension processes, but it is not certain in the current study how well L2 listeners would have understood the speech had their increased target-talker entrainment not occurred. Wöstmann, Herrmann, Maess, and Obleser (2016) found that listeners' performance on stimuli recall during dichotic listening was predicted by the hemispheric lateralisation of alpha (8-12 Hz) power, but not by low-frequency phase-locked responses to speech (also see Kerlin et al., 2010). Further studies are thus required to fully understand the effect of attention or listening effort on low-frequency entrainment to speech.

Korean listeners also had attenuated N400 effects compared to English listeners, which was driven by increased N400 amplitudes for words in the high cloze probability sentence condition. This demonstrates that non-native listeners need greater effort than do native listeners when processing predictable words, because predicting words using contextual cues is more difficult for non-native listeners due to their incomplete linguistic knowledge. This is consistent with Hahne (2001); non-native listeners exhibited a N400 response that was essentially similar to that of native listeners, but with quantitative differences in the semantically predicable condition. In addition to having less-developed semantic integration processes, it is possible that non-native

listeners' word recognition system was overwhelmed because they had to divert a greater amount of attentional resources to decoding the acoustic signal in this listening condition while maintaining speech segregation (i.e., a shortage of cognitive resources). As a result, they may have been less able to predict upcoming words using contextual cues. Similarly, previous research has suggested that the ability to exploit contextual information during speech processing is vulnerable to cognitive stress such as high working memory load (D'Arcy et al., 2005; Gunter et al., 1995). Speech degradation can have a similar consequence by limiting the availability of contextual cues (e.g., Aydelott et al., 2006; Obleser & Kotz, 2011), although the detrimental effect of speech degradation is more perceptual than cognitive.

In contrast, English listeners found this speech recognition task relatively easy; they thus had similar neural entrainment to the amplitude envelopes of target and distracting talkers. That is, they did not need to deploy additional attentional resources to selectively listen to the target talkers, which may be partly because the two talkers were presented to separate ears unlike some of the previous studies (e.g., Ding & Simon, 2012a). However, English listeners found Korean-accented English more difficult to understand than Standard Southern British English, as shown by the behavioural results. As a result, they had increased N400 amplitudes for the Korean-accented speech overall, indicating that they needed to exert greater effort to process words spoken with the less-intelligible L2 accent. This is partially consistent with the finding of Romero-Rivas et al. (2015) which found increased N400s for foreign-accented speech (i.e., only for semantic violations). Despite the relative difficulty in understanding Korean-accented speech, the magnitude of their N400 effect (i.e.,

difference between high- and low-probability conditions) was equal for both accents in the current study. That is, English listeners were still able to use contextual cues to predict upcoming words when listening to Korean-accented speech, demonstrating that the L1 speech recognition system is more robust to adverse conditions. This is also supported by the fact that they still outperformed Korean listeners on the recognition of Korean-accented speech in the behavioural task.

Taken together, the N400 results of the present study help resolve some of the inconsistencies between previous N400 studies. It seems that listeners can exert additional processing effort (i.e., larger N400) to help overcome some recognition problems, as did the native listeners when listening to L2-accented speech in the current study. However, the magnitude of the N400 effect can be reduced when speech intelligibility is severely decreased, whether due to acoustic degradation, high cognitive load or listener limitations (e.g., L2 or hearing-impaired listeners), thus making it difficult for listeners to exploit contextual cues (e.g., Aydelott et al., 2006; Obleser & Kotz, 2011).

Furthermore, Korean listeners had similar N400 responses for Southern British and Korean-accented English, in contrast to English listeners who exhibited greater N400 amplitudes for Korean-accented English. This result closely matches the pattern of accent intelligibility found in the behavioural responses of the present study as well as Study 1, with English listeners finding Korean-accented English less intelligible than their own accent, and Korean listeners finding both accents equally intelligible. This indicates that N400 is more closely related to speech intelligibility than is entrainment,

which was also supported by the positive correlation between individual listeners' N400 effect and behavioural accuracy. Moreover, the present work is the first to show that talker-listener accent interactions that have previously been found to affect accent intelligibility in behavioural studies (e.g., Bent & Bradlow, 2003; Pinet et al., 2011; see Chapter 2.1.4 for details) are also seen in neural processing of speech as reflected by the N400.

One possible limitation of the present study might be that there was no baseline condition with a single talker that could have isolated the effect of competing-talker background noise. That is, it would be of further interest to investigate the extent to which the observed differences between L1 and L2 listeners were caused by L2 listeners' general processing deficits (i.e., insufficient linguistic knowledge) or the vulnerability of their speech recognition system to the cognitive demands of the listening situation (i.e., competing-talker environments). Nonetheless, it should be noted that these two causes are not easily separable; the greater vulnerability to cognitive load during L2 speech perception also stems from having incomplete linguistic knowledge to a large extent. The current speech entrainment results also provided some evidence that L2 listeners indeed needed greater focused attention on the target speech signal than did L1 listeners. The observed differences between L1 and L2 listeners in the current study thus at least demonstrate that speech recognition was more effortful for L2 listeners than for L1 listeners.

Furthermore, a reliable correlation was not found between higher-level linguistic measures (i.e., N400 and behavioural response) and selective cortical entrainment to

target talkers within each listener group. The correlations might have been obscured because the inter-subject variability for these measures was small within each listener group. That is, each listener group was highly homogenous in terms of their English language competence. In future studies, it would therefore be interesting to test a correlation between target-speaker entrainment and other behavioural measures that allow for greater variation among listeners (e.g., listeners' subjective rating of listening effort, or performance on some attentional tasks), to further investigate the link between cortical entrainment to the speech envelope and listening effort. Future work also needs to determine whether the effect of top-down attention on speech entrainment occurs in a single-talker situation in response to other adverse conditions, or whether the increase in entrainment only reflects greater focused attention that is required to segregate a target speech stream from distracting speech streams or other background noise. In addition, the topographic distributions of target-speaker entrainment suggested that English listeners had a bilateral fronto-central distribution, whereas Korean listeners had a somewhat left-lateralised distribution. In future studies, it would be interesting to investigate what caused these differential distributions and where the increased entrainment of Korean listeners originated in the brain, using appropriate source analyses.

Chapter 5 General Discussion

Previous L2 speech research has mainly focused on the interaction of L1 and L2 phonologies in optimal laboratory conditions (both in terms of the listening condition and style of speech). However, the speech used in realistic communicative situations often deviates from canonical pronunciations that L2 listeners normally hear in the classroom, as it can be produced in a range of non-native and regional accents or in a casual speaking style. Furthermore, we must commonly comprehend speech in noise, which makes speech recognition more challenging both perceptually and cognitively. Although it has long been known that L2 listeners need to "listen harder", and even more so in these adverse conditions, it is not well-understood how these real-life factors affect L2 speech perception. This thesis explored these issues by examining how speech processing by L1 (native English) and L2 (native Korean) listeners is affected by speech style (read vs. casual speech), spoken accent, and adverse listening conditions (a competing-talker background), using electrophysiological and behavioural methods.

Study 1 was a behavioural speech-in-noise recognition study that investigated L2 speech recognition difficulties caused by casual speech. The results demonstrated that the detrimental effect of casual speech was greater for L2 listeners than for L1 listeners. Moreover, this study found talker-listener accent interactions regardless of speech style (read vs. casual speech), with English listeners displaying a clear advantage for their own accent (i.e., Standard Southern British English), and Korean listeners finding all accents similarly intelligible. Overall intelligibility differences between accents (i.e., Korean-accented English was less intelligible than SSBE or

Finnish-accented English) also decreased in the casual speech condition, indicating that certain characteristics of casual speech (e.g., less-reduced word forms, slower speaking rate) produced by inexperienced non-native speakers can be beneficial for the listener.

Study 2 was an EEG study which examined neural phase entrainment to the amplitude envelope of speech when subjects listened to their native language, second language or a language that they did not understand. The original aim of this study was to explore EEG methods that can measure how L1 and L2 listeners process more naturalistic, connected speech, but the cross-linguistic design of this experiment also allowed for an investigation of links between speech intelligibility and entrainment without altering the acoustic signals. Contrary to most of the previous findings, this study demonstrated that neural entrainment to speech is purely modulated by acoustic properties of the speech signals rather than the listener's higher-level linguistic processing.

Study 3 investigated speech recognition in a competing-talker background using entrainment and N400 measures as well as a behavioural speech recognition task. This study produced a surprising result; L2 listeners had more selective auditory processing (i.e., higher entrainment) for target talkers than did L1 listeners likely because they needed greater listening effort to process the target signal. L2 listeners also had greater lexical processing (i.e., larger N400) for words in highly predictable contexts than did L1 listeners. In contrast, L1 listeners increased their degree of lexical processing in

response to Korean-accented speech. Overall, these results show different ways of adapting to adverse conditions.

The methodological goal of this thesis was to develop new tools that are suitable for investigating these L2 speech recognition problems encountered in real-life situations. Specifically, the use of spontaneous speech has been highly limited in speech perception research (e.g., shorter fragments of speech have been mostly used), because of the methodological challenge caused by its highly variable and uncontrolled nature. Study 1 presented a new method of evaluating speech recognition performance for spontaneous speech (i.e., picture evaluation task). This task appears to be highly efficient as it does not require listeners to repeat back or write down spontaneous utterances that can be relatively unstructured. Although the identification of each keyword cannot be measured with this method, a larger number of trials can be tested instead, given that subjects spend much less time giving a respond than in typical speech recognition tasks. This method has great potential for future research in evaluating speech recognition performance for spontaneous speech materials or for populations who may need a task that is more enjoyable and visually attractive such as children.

In addition, the original aim of Study 2 was to explore EEG measures that can assess how L2 listeners process continuous speech rather than focusing on the perception of isolated syllables or phonemes. Contrary to expectations, the results suggested that this neural measure was not sensitive to listeners' language experience or higher-level linguistic processing (i.e., comprehension). Instead, in two-talker situations examined

in Study 3, the additional listening effort experienced by L2 listeners was found to enhance their neural tracking of target talkers, suggesting a link between neural entrainment and listening effort. Furthermore, Study 3 has shown that the EEG paradigm used in this study is able to examine speech processing at both auditory and lexical levels within a single experiment using measures of cortical entrainment and N400. This method can allow for a wide range of investigations of L2 speech recognition problems that occur at multiple levels (i.e., auditory and higher linguistic), and also has broad relevance for other populations such as hearing-impaired or older listeners.

The results of the current thesis also point to interesting new directions for future work. Investigating what features of native- and foreign-accented casual speech enhance or degrade speech intelligibility can be a direction of future L2 research; this question can be addressed by focusing on the effect of each of the relevant acoustic-phonetic and phonological features (e.g., vowel reduction/deletion or reduced vowel space). In order to compare spontaneous speech with read speech in future research, it would also be advantageous to explore other methods to have completely matched materials for read and casual speech; for example, one can have speakers read the utterances that they have spontaneously produced (e.g., Mehta & Cutler, 1988).

As mentioned above, the results of Study 2 demonstrated that listeners had similar entrainment for all languages. This seems to suggest that listeners' L1 experience does not necessarily alter how they process speech at the cortical level, as far as speech entrainment is concerned. However, it should be noted that the amplitude envelope of

speech does not necessarily provide clear syllable boundaries (Cummins, 2012), which suggests that the effect of L1 experience might be found if phase entrainment is measured for specific syllable types (e.g., CCV). For example, previous research has suggested that L1 phonotactic constraints play a role in perception; for example, Japanese listeners have been reported to hear an illusionary vowel 'u' between the two consonants in VCCV stimuli because their native phonology does not allow complex consonant clusters (e.g., Dupoux, Hirose, Kakehi, Pallier, & Mehler, 1999). If cross-linguistic differences are found in entrainment to syllables, this may help elucidate the underlying mechanisms of 'syllable parsing' that was proposed in previous studies (Giraud & Poeppel, 2012; Ghitza, 2013), and its associations with listeners' phonological knowledge.

It would also be interesting to further investigate the link between entrainment and listening effort; the present work could be extended to other listener populations (e.g., hearing-impaired listeners) or other listening conditions (e.g., single-talker environments). In addition, a more thorough investigation could be conducted using stimuli of varying levels of intelligibility (e.g., different target-to-masker ratios), in order to determine how target-talker entrainment is modulated by speech intelligibility and perception difficulty. Furthermore, it remains for future research to determine whether the enhanced entrainment of L2 listeners facilitates speech comprehension. To understand the nature of this enhancement, it would also be interesting to examine exactly where this increased phase-locking occurs in the brain using methods such as MEG. For example, it could occur in certain brain regions that are activated during effortful listening (e.g., anterior cingulate cortex).

In sum, the current thesis demonstrated some important factors that can account for L2 speech perception difficulties in everyday speech communication. Specifically, casual speech that occurs in everyday situations poses additional challenges for L2 listeners. Listeners also need to "listen harder", especially in challenging situations such as in a competing-talker background, but this increased listening effort experienced by L2 listeners can enhance their neural tracking of attended speech streams. This may show a compensatory mechanism that L2 listeners deploy to overcome their perceptual challenges under adverse conditions. This result also suggests that cognitive factors (e.g., cognitive load) play a more important role in L2 speech perception than previously thought, by altering specific aspects of speech perception. In contrast, the present work suggests that native listeners are generally less affected by these adverse conditions, and are able to flexibly modulate their processing to fit the demands of the listening condition (e.g., a difficult L2 accent) by deploying additional effort at the level of lexical processes. It thus appears that adverse conditions can have diverse effects on the auditory and lexical processing of speech, and that these effects differ for L1 and L2 listeners. These lines of research will be interesting avenues for future L2 research.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(2), 520–529. https://doi.org/10.1037/a0013552

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, a, Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–13372. https://doi.org/10.1073/pnas.201400998

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for C ontinuous Mapping Models. *Journal of Memory and Language*, *38*(38), 419–439. https://doi.org/10.1006/jmla.1997.2558

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., … Weinert, R. (1991). the Hcrc Map Task Corpus. *Language and Speech*, *34*(4), 351–366. https://doi.org/10.1177/002383099103400404

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The Effects of Nonnative Accents on Listening Comprehension: Implications for ESL. *Source: TESOL Quarterly*, *36*(2), 173–190. https://doi.org/10.2307/3588329

Aydelott, J., Dick, F., & Mills, D. L. (2006). Effects of acoustic distortion and semantic context on event-related potentials to spoken words. *Psychophysiology*, *43*(5), 454–464. https://doi.org/10.1111/j.1469-8986.2006.00448.x

Baker, R., & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*(3), 761–770. https://doi.org/10.3758/s13428-011-0075-y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 51. https://doi.org/10.18637/jss.v067.i01

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610. https://doi.org/10.1121/1.1603234

Best, C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 171–204). Timonium, MD: York Press.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794. https://doi.org/10.1121/1.1332378

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, *84*(1), 101–114. https://doi.org/10.1121/1.396976

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat:" the effects of semantic and acoustic information on phoneme categorization. *The Journal of the Acoustical Society of America*, *103*(5), 2670–2676. https://doi.org/10.1121/1.422787

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer (Version 5.3.69) [Computer program]. Retrieved from http://www.praat.org/

Borges, J. L. (1949). La casa de Asterión [The House of Asterion]. In Borges, J. L. (Ed), *The Aleph.* Retrieved from http://ciudadseva.com/texto/la-casa-de-asterion/

Boulenger, V., Hoen, M., Jacquier, C., & Meunier, F. (2011). Interplay between acoustic/phonetic and semantic processes during spoken sentence comprehension: An ERP study. *Brain and Language*, *116*(2), 51–63. https://doi.org/10.1016/j.bandl.2010.09.011

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Bradlow, A. R., & Alexander, J. a. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349. https://doi.org/10.1121/1.2642103

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, *112*(1), 272–284. https://doi.org/10.1121/1.1487837

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Bradlow, A. R., Kraus, N., & Hayers, E. (2003). Speaking clearly for children with learning disabilities: Sentence perception in noise. *Journal of Speech, Language, and Hearing Research, 48*, 80-97. http://dx.doi.org/10.1121/1.4743692

Brandmeyer, A., Desain, P. W. M., & McQueen, J. M. (2012). Effects of native language on perceptual sensitivity to phonetic cues. *NeuroReport*, *23*, 653–657. https://doi.org/10.1097/WNR.0b013e32835542cd

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539–571. https://doi.org/10.1080/01690965.2011.555268

Brown, C., & Hagoort, P. (1993). The Processing Nature of the N400: Evidence from Masked Priming. *Journal of Cognitive Neuroscience*, *5*(1), 34–44. https://doi.org/10.1162/jocn.1993.5.1.34

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers, *The Journal of the Acoustical Society of America, 109*, 1101–1109. http://dx.doi. org/10.1121/1.1345696.

Brungart, D. S. (2005). Informational and energetic masking effects in multitalker speech perception. *Speech Separation by Humans and Machines*, *1101*(2001), 261–267. https://doi.org/10.1007/0-387-22794-6_17

Burnett, F. H. (1909). *The Secret Garden*. London, Portsmouth: Heineman. Retrieved from http://etc.usf.edu/lit2go/163/the-secret-garden/

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Calandruccio, L., & Smiljanic, R. (2012). New Sentence Recognition Materials

Developed Using a Basic Non-Native English Lexicon. *Journal of Speech, Language, and Hearing Research*, *55*(5), 1342–1355. https://doi.org/10.1044/1092-4388(2012/11-0260)

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *NeuroImage*, *22*(3), 1182–1194. https://doi.org/10.1016/j.neuroimage.2004.03.006

Campbell, J., & Sharma, A. (2013). Compensatory changes in cortical resource allocation in adults with hearing loss. *Frontiers in Systems Neuroscience*, *7*, 1-9. https://doi.org/10.3389/fnsys.2013.00071

Carey, D., Mercure, E., Pizzioli, F., & Aydelott, J. (2014). Auditory semantic processing in dichotic listening: Effects of competing speech, ear of presentation, and sentential bias on N400s to spoken words in context. *Neuropsychologia*, *65*, 102–112. https://doi.org/10.1016/j.neuropsychologia.2014.10.016

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? *Trends in Cognitive Sciences*, *10(12)*, 564–570. https://doi.org/10.1016/j.tics.2006.10.002

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 283-333). Cambridge: Cambridge University Press.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language*, *26*(5), 527–538. https://doi.org/10.1016/0749-596X(87)90138-0

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*(1), 414–427. https://doi.org/10.1121/1.2804952

Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *The Journal of the Acoustical Society of America*, *128*(4), 2059–2069. https://doi.org/10.1121/1.3478775

Cortázar, J. (1947). Casa tomada [House taken over]. In J. L. Borges (Ed), *Los anales de Buenos Aires.* Retrieved from http://ciudadseva.com/texto/casa-tomada/

Crawford, M. D., Brown, G. J., Cooke, M. P., & Green, P. D. (1994). The design, collection and annotation of a multi-agent, multi- sensor speech corpus. *Proceedings of the Institute of Acoustics, 16,* 183–189.

Cummins, F. (2012). Oscillators and syllables: A cautionary note. *Frontiers in Psychology*, *3*, 1–2. https://doi.org/10.3389/fpsyg.2012.00364

Cummins, F., Grimaldi, M., Leonard, T., & Simko, J. (2006). The CHAINS Speech Corpus: CHAracterizing INdividual Speakers. *Proceedings of SPECOM*, 1–6.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, *24*(3), 381–410. https://doi.org/10.1016/0010-0285(92)90012-Q

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121. https://doi.org/10.1037/0096-1523.14.1.113

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284. https://doi.org/10.1016/j.wocn.2005.06.002

D'Arcy, R. C. N., Service, E., Connolly, J. F., & Hawco, C. S. (2005). The influence of increased working memory load on semantic neural systems: A high-resolution event-related brain potential study. *Cognitive Brain Research*, *22*(2), 177–191. https://doi.org/10.1016/j.cogbrainres.2004.08.007

Darcy, I., Ramus, F., Christophe, A., Kinzler, K., & Dupoux, E. (2009). Phonological knowledge in compensation for native and non-native assimilation. In F. Kügler, C. Féry, & R. van de Vijver (Eds.), *Variation and Gradience in Phonetics and Phonology* (pp. 265–310). Berlin: Mouton De Gruyter.

Darwin, C. J., & Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America*, *107*(2), 970–977. https://doi.org/10.1121/1.428278

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *23*(8), 3423–3431. https://doi.org/23/8/3423 [pii]

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147. https://doi.org/10.1016/j.heares.2007.01.014

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222

De Cheveigné, A., & Parra, L. C. (2014). Joint decorrelation, a versatile tool for multichannel data analysis. *NeuroImage*, *98*, 487–505. https://doi.org/10.1016/j.neuroimage.2014.05.068

de Cheveigné, A., & Simon, J. Z. (2008). Denoising based on spatial filtering. *Journal of Neuroscience Methods*, *171*(2), 331–339. https://doi.org/10.1016/j.jneumeth.2008.03.015

Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., … Le Bihan, D. (1997). Anatomical variability in the cortical representation of first and second language. *Neuroreport*, *8*(17), 3809–3815.

https://doi.org/10.1097/00001756-199712010-00030

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *Neuroreport*, *8*(4), 919–924. https://doi.org/10.1097/00001756-199703030-00021

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Ding, N., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, *88*, 41–46. https://doi.org/10.1016/j.neuroimage.2013.10.054

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–64. https://doi.org/10.1038/nn.4186

Ding, N., & Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11854–9. https://doi.org/10.1073/pnas.1205381109

Ding, N., & Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

Ding, N., & Simon, J. Z. (2013). Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. *Journal of Neuroscience Current Issue*, *33*(13), 5728–5735. https://doi.org/10.1523/JNEUROSCI.5297-12.2013

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768.

https://doi.org/10.1016/j.neuroimage.2013.06.035

Dupoux, E., Hirose, Y., Kakehi, K., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology Human Perception and Performance*, *25*(6), 1568–1578. https://doi.org/10.1037/0096-1523.25.6.1568

Eckert, M. A., Menon, V., Walczak, A., Ahlstrom, J., Denslow, S., Horwitz, A., & Dubno, J. R. (2009). At the heart of the ventral attention system: The right anterior insula. *Human Brain Mapping*, *30*(8), 2530–2541. https://doi.org/10.1002/hbm.20688

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *Journal of Neuroscience*, *33*(26), 10688–97. https://doi.org/10.1523/JNEUROSCI.4596-12.2013

Erb, J., & Obleser, J. (2013). Upregulation of cognitive control networks in older adults' speech comprehension. *Frontiers in Systems Neuroscience*, *7*, 116. https://doi.org/10.3389/fnsys.2013.00116

Ernestus, M., Baayen, R. H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, *81*(1–3), 162–173. https://doi.org/10.1006/brln.2001.2514

Evans, B. G., & Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *The Journal of the Acoustical Society of America*, *121*(6), 3814. https://doi.org/10.1121/1.2722209

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *112*(1), 259–271. https://doi.org/10.1121/1.1482078

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research : JSLHR*, *50*(5), 1241–1255.

https://doi.org/10.1044/1092-4388(2007/087)

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America, 93*(3), 1589–1608. https://doi.org/10.1121/1.406818

Flege, J. E. (1995). Second Language Speech Learning: Theory, Findings, and Problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research,* 233–277. https://doi.org/10.1111/j.1600-0404.1995.tb01710.x

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Flege, J. E., & Davidian, R. D. (1984). Transfer and developmental processes in adult foreign language speech production. *Applied Psycholinguistics, 5,* 323–347. https://doi.org/10.1017/S014271640000521X

Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America, 106*(5), 2973–2987. https://doi.org/10.1121/1.428116

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America, 97*(5), 3125–3134. https://doi.org/10.1121/1.413041

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age Constraints on Second-Language Acquisition. *Journal of Memory and Language, 41*(1), 78–104. https://doi.org/10.1006/jmla.1999.2638

Fosler-Lussier, E., & Morgan, N. (1999). Effects of speaking rate and word frequency on pronunciation in conversational speech. *Speech Communication, 29,* 137–158. https://doi.org/10.1016/S0167-6393(99)00035-7

Fox, J., & Weisberg, S. (2011). *An {R} Companion to Applied Regression*. Thousand Oaks, CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech

recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256. https://doi.org/10.1121/1.1689343

Gaskell, M. G., & Marslen-Wilson, W. D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(1), 144–58. https://doi.org/10.1037/0096-1523.22.1.144

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*, 1–13. https://doi.org/10.3389/fpsyg.2011.00130

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, *4*, 1–5. https://doi.org/10.3389/fpsyg.2013.00138

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1–2), 113–126. https://doi.org/10.1159/000208934

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–7. https://doi.org/10.1038/nn.3063

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD telephone speech corpus for research and development. *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92), 1*, 517–520. https://doi.org/10.1109/ICASSP.1992.225858

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language*, *122*(2), 92–102. https://doi.org/10.1016/j.bandl.2012.04.017

Gow, D. W. (2002). Does English coronal place assimilation create lexical ambiguity? *Journal of Experimental Psychology: Human Perception and*

Performance, *28*(1), 163–179. https://doi.org/10.1037//0096-1523.28.1.163

Grabe, E., Post, B., & Nolan, F. (2001). The IViE Corpus. Department of Linguistics, University of Cambridge. Retrieved from http://www.phon.ox.ac.uk/IViE

Greenberg, S. (1999). Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*(2), 159–176. https://doi.org/10.1016/S0167-6393(99)00050-3

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12), 1-14. https://doi.org/10.1371/journal.pbio.1001752

Gunter, T. C., Jackson, J. L., & Mulder, G. (1995). Language, memory, and aging: An electrophysiological exploration of the N400 during reading of memory demanding sentences. *Psychophysiology*, *32*(3), 215–229. https://doi.org/10.1111/j.1469-8986.1995.tb02951.x

Guy, G. R. (1980). Variation in the group and the individual. In W. Labov (Ed.), *Locating language in time and space* (pp. 1–36). New York, NY: Academic Press.

Hagoort, P. (2008). The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1055–69. https://doi.org/10.1098/rstb.2007.2159

Hahne, A. (2001). What ' s Different in Second-Language Processing ? Evidence from Event-Related. *Journal of Psycholinguistic Research*, *30*(3), 251–266. https://doi.org/10.1023/A:1010490917575

Hahne, A., & Friederici, A. D. (2001). Processing a second language: Late learners' comprehension mechanisms as revealed by event-related brain potentials. *Bilingualism: Language and Cognition*, *4*(2), 123–141. https://doi.org/10.1017/S1366728901000232

Hald, L. A., Bastiaansen, M. C. M., & Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain and*

Language, *96*(1), 90–105. https://doi.org/10.1016/j.bandl.2005.06.007

Halliwell-Phillipps. J. O. (n.d.), *Lazy Jack*, Retrieved from http://www.storynory.com/

Hambrook, D. A., & Tata, M. S. (2014). Theta-band phase tracking in the two-talker problem. *Brain and Language*, *135*, 52–56. https://doi.org/10.1016/j.bandl.2014.05.003

Hanulíková, A., van Alphen, P. M., van Goch, M. M., & Weber, A. (2012). When One Person's Mistake Is Another's Standard Usage: The Effect of Foreign Accent on Syntactic Processing. *Journal of Cognitive Neuroscience*, *24*(4), 878–887. https://doi.org/10.1162/jocn_a_00103

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America*, *125*(1), 469–479. https://doi.org/10.1121/1.3021295

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3–4), 373–405. https://doi.org/10.1016/j.wocn.2003.09.006

Haynes, R. M., White, L., & Mattys, S. L. (2015). What do we expect spontaneous speech to sound like? *ICPhS 2015. Proceedings of the 18th International Congress of Phonetic Sciences*. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1011.pdf

Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *The Journal of the Acoustical Society of America*, *130*(4), 2139. https://doi.org/10.1121/1.3623753

Hazan, V., Grynpas, J., & Baker, R. (2012). Is clear speech tailored to counter the effect of specific adverse listening conditions? *The Journal of the Acoustical Society of America*, *132*(5), EL371. https://doi.org/10.1121/1.4757698

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *Journal of the Acoustical Society of*

America, *116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

Hazan, V., & Simpson, A. (2000). The Effect of Cue-Enhancement on Consonant Intelligibility in Noise: Speaker and Listener Effects. *Language and Speech*, *43*(3), 273–294. https://doi.org/10.1177/00238309000430030301

Heungbuwa Nolbu [Heungbu and Nolbu] (2003). Retrieved from http://terms.naver.com/

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Horton, C., D'Zmura, M., & Srinivasan, R. (2013). Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology*, *109*(12), 3082–3093. https://doi.org/10.1152/jn.01026.2012

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal 50*(3), 346-363.

Howard, M. F., & Poeppel, D. (2010). Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology*, *104*, 2500–2511. https://doi.org/10.1152/jn.00251.2010

HTK Hidden Markov Modelling toolkit [Computer software] (1989). Available from http://htk.eng.cam.ac.uk/

Huckvale, M. (2004). ACCDIST: A metric for comparing speakers' accents. *International Conference on Spoken Language Processing (INTERSPEECH)*, 1–4. Retrieved from http://discovery.ucl.ac.uk/12139/

Huckvale, M. (2007a). Hierarchical clustering of speakers into accents with the ACCDIST metric. *International Conference of Phonetic Sciences (ICPhS)*, 1–4.

Huckvale, M. (2007b). ACCDIST: an accent similarity metric for accent recognition and diagnosis. In C. Müller (Ed.), *Speaker Classification II* (pp. 258-275). Berlin: Springer.

Imai, S., Walley, A. C., & Flege, J. E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by

native English and Spanish listeners. *The Journal of the Acoustical Society of America*, *117*(2), 896–907. https://doi.org/10.1121/1.1823291

Indefrey, P. (2006). A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean? *Language Learning*, *56*, 279–304. https://doi.org/10.1111/j.1467-9922.2006.00365.x

Ito, A., Corley, M., & Pickering, M. J. (2017). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension, *Bilingualism: Language and Cognition,* 1-14. https://doi.org/10.1017/S1366728917000050

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. doi:10.1016/S0

Iverson, P., Pinet, M., & Evans, B. G. (2014). Mapping accent similarity and speech in noise intelligibility for British English accents. *The Journal of the Acoustical Society of America, 135,* 2422. http://dx.doi.org/10.1121/1.4878048

Johnson, K. (1997). Speech perception without speaker normalization: an exemplar model. In K. John-son & J. Mullenix (Eds.), *Talker variability in speech processing* (pp.145–166). San Diego, CA: Academic Press.

Johnson, K. (2004). Massive reduction in conversational American English. *Proceedings of the Workshop on Spontaneous Speech: Data and Analysis*, 29–54. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=77FFE3C22602F55D083CAB7CA669DB0E?doi=10.1.1.142.5012&rep=rep1&type=pdf

Jurafsky, D., Bell, A., Gregory, M., & Raymond, W. D. (2000). Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. In J. Bybee and P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 229–254). Amsterdam: John Benjamins.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word

predictability. *The Journal of the Acoustical Society of America*, *61*(5), 1337–1351. https://doi.org/10.1121/1.381436

Kemps, R., Ernestus, M., Schreuder, R., & Baayen, H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, *90*(1–3), 117–127. https://doi.org/10.1016/S0093-934X(03)00425-5

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a “cocktail party”. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(2), 620–628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010

Kimura, D. (1961). Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psycholology*, *15*, 166–171. http://dx.doi.org/10.1037/h0083219

Kong, Y.-Y., Somarowthu, A., & Ding, N. (2015). Effects of Spectral Degradation on Attentional Modulation of Cortical Auditory Responses to Continuous Speech. *Journal of the Association for Research in Otolaryngology*, *16*(6), 783–796. https://doi.org/10.1007/s10162-015-0540-x

Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, *115*(1), 362–378. https://doi.org/10.1121/1.416659

Krishnan, A., Gandour, J. T., Bidelman, G. M., & Swaminathan, J. (2009). Experience-dependent neural representation of dynamic pitch in the brainstem. *NeuroReport*, *20*(4), 408–413. https://doi.org/10.1097/WNR.0b013e3283263000

Krishnan, A., Swaminathan, J., & Gandour, J. T. (2008). Experience-dependent Enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *Journal of Cognitive Neuroscience*, *21*(6), 1092–1105. https://doi.org/10.1162/jocn.2009.21077

Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, *25*(1), 161–168. https://doi.org/10.1016/j.cogbrainres.2005.05.004

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current*

Opinion in Neurobiology, *4*(6), 812–822. https://doi.org/10.1016/0959-4388(94)90128-7

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2). https://doi.org/10.1111/j.1467-7687.2006.00468.x

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608. https://doi.org/10.1126/science.1736364

Kutas, M., & Federmeier, K. D. (2000). Electropsysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Science*, *12*(12), 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 14.1–14.27. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207(4427)*, 203-205. https://doi.org/10.1126/science.7350657

Kwon, J. (2007). *Mongsil Unni* [Sister Mongsil]. Seoul: Changbi Publishers.

Lachaux, J. P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping*, *8*(4), 194–208. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C

Lahiri, A., & Marslen-Wilson, W.D. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. *Cognition, 38*, 245-294. doi: 10.1016/0010-0277(91)90008-R

Lahiri, A., & Marslen-Wilson, W.D. (1992). Lexical processing and phonological representations. In G.J. Docherty & D.R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 229-254). Cambridge: Cambridge University Press.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. *Journal of Neurophysiology*, *94*, 1904–1911. https://doi.org/10.1152/jn.00263.2005

Lalor, E. C., Power, A. J., Reilly, R. B., & Foxe, J. J. (2009). Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. *Journal of Neurophysiology*, *102*(1), 349–359. https://doi.org/10.1152/jn.90896.2008

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/Doi 10.1038/Nrn2532

Lecumberri, M. L. G., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, *119*(4), 2445–2454. https://doi.org/10.1121/1.2180210

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Lee, H. B., Jin, N., Seong, C., Jung, I., & Lee, S. (1994). An experimental phonetic study of speech rhythm in Standard Korean. *Proceedings of International Conference on Spoken Language Processing*, 1091–1094.

Lemke, U., & Besser, J. (2016). Cognitive Load and Listening Effort : Concepts and Age-Related Considerations. *Ear & Hearing*, *37*, 77S–84S. https://doi.org/10.1097/AUD.0000000000000304

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In: W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403-439). Dordrecht: Kluwer Academic Publishers.

Liu, S., & Zeng, F.-G. (2006). Temporal properties in clear speech perception. *The Journal of the Acoustical Society of America*, *120*(1), 424–432. https://doi.org/10.1121/1.2208427

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 1–14. https://doi.org/10.3389/fnhum.2014.0021

Luck, S. J. (2005). *An introduction to the event-related potential technique.* Cambridge, MA: The MIT Press.

Luck, S., & Kappenman, E. (2012). *The Oxford handbook of event-related potential components.* Oxford: Oxford University Press.

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*(1–2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin & Review*, *21*(3), 748–754. https://doi.org/10.3758/s13423-013-0544-7

Mattys, S. L., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*(3), 203–243. https://doi.org/10.1016/j.cogpsych.2009.04.001

Mattys, S. L., Carroll, L. M., Li, C. K. W., & Chan, S. L. Y. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech Communication*, *52*(11–12), 887–899. https://doi.org/10.1016/j.specom.2010.01.005

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions : A review. *Language and Cognitive processes*, 27(7/8), 953-978. http://dx.doi.org/10.1080/01690965.2012.705006

Mattys, S. L., & Palmer, S. D. (2015). Divided attention disrupts perceptual encoding during speech recognition. *The Journal of the Acoustical Society of America*,

137(3), 1464–1472. https://doi.org/10.1121/1.4913507

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of Multiple Speech Segmentation Cues: A Hierarchical Framework. *Journal of Experimental Psychology: General*, *134*(4), 477–500. https://doi.org/10.1037/0096-3445.134.4.477

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*(2), 145–160. https://doi.org/10.1016/j.jml.2011.04.004

Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of Second-Language Acquisition and Perception of Speech in Noise. *Journal of Speech Language and Hearing Research*, *40*(3), 686. https://doi.org/10.1044/jslhr.4003.686

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: what exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group “white paper”. *International Journal of Audiology*, *53*(7), 433–40. https://doi.org/10.3109/14992027.2014.890296

McQueen, J. M. (1998). Segmentation of Continuous Speech Using Phonotactics. *Journal of Memory and Language*, *39*(1), 21–46. https://doi.org/10.1006/jmla.1998.2568

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*(1), 509. https://doi.org/10.1121/1.3664087

Mehta, G., & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech, 31*(2), 135-156.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The Intelligibility of Speech As a Function of the Context of the Test Materials. *Journal of Experimental Psychology*, *41*(5), 329–335. https://doi.org/10.1023/A

Millman, R. E., Johnson, S.R., & Prendergast, G. (2015). The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. *Journal of Cognitive Neuroscience, 27*(3), 533-45. doi: 10.1162/jocn_a_00719

Mitterer, H., & Tuinman, A. (2012). The role of native-language knowledge in the perception of casual speech in a second language. *Frontiers in Psychology*, *3*, 1–13. https://doi.org/10.3389/fpsyg.2012.00249

Molloy, K., Griffiths, T. D., Chait, M., & Lavie, N. (2015). Behavioral/Cognitive Inattentional Deafness: Visual Load Leads to Time-Specific Suppression of Auditory Evoked Responses. *The Journal of Neuroscience*, *35*(49), 16046–16054. https://doi.org/10.1523/JNEUROSCI.2931-15.2015

Monterroso, A. (1958). *El eclipse* [The eclipse]. Retrieved from http://ciudadseva.com/texto/el-eclipse/

Moon, S.-J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *The Journal of the Acoustical Society of America*, *96*(1), 40-55. https://doi.org/10.1121/1.410492

Mueller, J. L., Hahne, A., Fujii, Y., & Friederici, A. D. (2005). Native and nonnative speakers' processing of a miniature version of Japanese as revealed by ERPs. *Journal of Cognitive Neuroscience*, *17*(8), 1229–1244. https://doi.org/10.1162/0898929055002463

Mueller, J. L., Planck, M., & Cognitive, H. (2005). Electrophysiological correlates of second language processing, *Second Language Research, 21(2),*152–174. https://doi.org/10.1191/0267658305sr256oa

Munro, M. (1998). The effects of noise on the intelligibility of foreign- accented speech, *Studies of Second Language Acquisition, 20*, 139–154. https://doi.org/10.1017/S0272263198002022

Munro, M., & Derwing, T. (1995). Processing time, accent and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38(3),* 289–306. doi:10.1177/002383099503800305

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Livonen, A., . . . Alho, K. (1997). Language-Specific Phoneme Representations Revealed by

Electric and Magnetic Brain Responses. *Nature, 385*(6615), 432-434. doi:10.1038/385432a0

Näätänen, Nábělek, A. K., & Donahue, A. M. (1984). Perception of consonants in reverberation by native and non-native listeners. *Journal of the Acoustical Society of America, 75*(2), 632–634. https://doi.org/10.1121/1.390495

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The Influence of Language Proficiency on Lexical Semantic Processing in Native and Late Learners of English. *Journal of Cognitive Neuroscience, 24*(5), 1205–1223. https://doi.org/10.1162/jocn_a_00143

Norris, D. (1994). Shortlist - a Connectionist Model of Continuous Speech Recognition. *Cognition, 52*(3), 189–234. https://doi.org/10.1016/0010-0277(94)90043-4

Norris, D., McQueen, J. M., & Cutler. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*(3), 299-325. https://doi.org/10.1017/S0140525X00003241

Norris, D., McQueen, J. M., Cutler, a, & Butterfield, S. (1997). The possible-word constraint in the segmentation of continuous speech. *Cognitive Psychology, 34*(3), 191–243. https://doi.org/10.1006/cogp.1997.0671

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., … Brugge, J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. *Journal of Neuroscience, 29*(49), 15564–15574. https://doi.org/10.1523/JNEUROSCI.3065-09.2009

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex, 25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Obleser, J., Herrmann, B., & Henry, M. J. (2012). Neural Oscillations in Speech: Don't be Enslaved by the Envelope. *Frontiers in Human Neuroscience, 6*, 1–4. https://doi.org/10.3389/fnhum.2012.00250

Obleser, J., & Kotz, S. A. (2010). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex, 20*(3), 633–

640. https://doi.org/10.1093/cercor/bhp128

Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *NeuroImage*, *55*(2), 713–723. https://doi.org/10.1016/j.neuroimage.2010.12.020

Obleser, J., Wise, R. J. S., Dresner, M. A., & Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, *27*(9), 2283–2289. https://doi.org/10.1523/JNEUROSCI.4663-06.2007

Obleser, J., Wostmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse Listening Conditions and Memory Load Drive a Common Alpha Oscillatory Network. *Journal of Neuroscience*, *32*(36), 12376–12383. https://doi.org/10.1523/JNEUROSCI.4908-11.2012

Oh, J. S., Jun, S. A., Knightly, L. M., & Au, T. K. F. (2003). Holding on to childhood language memory. *Cognition*, *86*(3), B53–B64. https://doi.org/10.1016/S0010-0277(02)00175-0

Oostdijk, N. (2000). The spoken Dutch corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer (Eds), *Proceedings of the Second International Conference on Language Resources and Evaluation*, (Vol. 2, pp. 887–893), Paris. ELRA.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*. 1-9. https://doi.org/10.1155/2011/156869

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of

conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, *95*(3), 1581–1592. https://doi.org/10.1121/1.408545

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*, 1–17. https://doi.org/10.3389/fpsyg.2012.00320

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Peelle, J. E., Troiani, V., Grossman, M., & Wingfield, A. (2011). Hearing loss in older adults affects neural systems supporting speech comprehension. *The Journal of Neuroscience*, *31*(35), 12638–12643. https://doi.org/10.1523/JNEUROSCI.2559-11.2011

Peña, M., & Melloni, L. (2012). Brain Oscillations during Spoken Sentence Processing. *Journal of Cognitive Neuroscience*, *24*(5), 1149–1164. https://doi.org/10.1162/jocn_a_00144

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking Clearly for the Hard of Hearing II. *Journal of Speech Language and Hearing Research*, *29*, 434-446. https://doi.org/10.1044/jshr.3203.600

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), Frequency and the emergence of linguistic structure (Vol. 45, pp. 137–157). Amsterdam: John Benjamins. https://doi.org/10.1.1.142.4394

Pinet, M., Iverson, P., & Huckvale, M. (2011). Second-language experience and speech-in-noise recognition: effects of talker-listener accent similarity. *The Journal of the Acoustical Society of America*, *130*(3), 1653–62. https://doi.org/10.1121/1.3613698

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd Release). Columbus, OH: Department of Psychology, Ohio State University (Distributor). Retreived from www.buckeyecorpus.osu.edu

Piquado, T., Benichov, J. I., Brownell, H., & Wingfield, A. (2012). The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening. *International Journal of Audiology, 51*(8), 576–83. https://doi.org/10.3109/14992027.2012.684403

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time." *Speech Communication, 41*(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Poeppel, D., Idsardi, W. J., & van Wassenhove, V. (2008). Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363*(1493), 1071–1086. https://doi.org/10.1098/rstb.2007.2160

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*(3), 265–292. https://doi.org/10.1016/S0010-0277(99)00058-X

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology, 9*(4), 1-15. https://doi.org/10.1371/journal.pbio.1000610

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *The Journal of the Acoustical Society of America, 118*(3), 1274–1277. https://doi.org/10.1121/1.2000751

Rimmele, J. M., Zion Golumbic, E., Schröger, E., & Poeppel, D. (2015). The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene. *Cortex, 68*, 144–154. https://doi.org/10.1016/j.cortex.2014.12.014

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., & Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics, 27*, 465–485. https://doi.org/10.1017.S014271640606036X

Romero-Rivas, C., Martin, C. D., & Costa, A. (2015). Processing changes when listening to foreign-accented speech. *Frontiers in Human Neuroscience, 9*, 1-15. https://doi.org/10.3389/fnhum.2015.00167

Rönnberg, J. (2003). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: a framework and a model. *International Journal of Audiology*, *42*, 68–76. https://doi.org/10.3109/14992020309074626

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *International Journal of Audiology*, *47*, S99–S105. https://doi.org/10.1080/14992020802301167

Rönnberg, J., Rudner, M., Lunner, T., & Zekveld, A. (2010). When cognition kicks in: working memory and speech understanding in noise. *Noise & Health*, *12(49)*, 263–269. https://doi.org/10.4103/1463-1741.70505

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, *336*(1278), 367–373.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, *398*(6730), 760. https://doi.org/10.1038/19652

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Frontiers in Psychology*, *5*, 1-16. https://doi.org/10.3389/fpsyg.2014.00137

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences*, *32*(1), 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Seong, C. J. (1995). The Experimental phonetic study of the standard current Korean speech rhythm: with respect to its temporal structure (Doctoral dissertation). Seoul National Univesity, Seoul.

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *The Journal of the Acoustical Society of America*, *107*(5),

2697–2703. https://doi.org/10.1121/1.428655

Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between N1 evoked potential morphology and the perception of voicing. *The Journal of the Acoustical Society of America*, *108*(6), 3030–5. https://doi.org/10.1121/1.1320474

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243. https://doi.org/10.1017/S0142716400001417

Smiljanic, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass, 3(1)*, 236–264. doi: 10.1111/j.1749-818X.2008.00112.x

Spivey, M. J., & Marian, V. (1999). Cross Talk Between Native and Second Languages: Partial Activation of an Irrelevant Lexicon. *Psychological Science*, *10*(3), 281–284. https://doi.org/10.1111/1467-9280.00151

Steinschneider, M., Nourski, K. V., & Fishman, Y. I. (2013). Representation of speech in human auditory cortex: Is it special? *Hearing Research*, *305*(1), 57–73. https://doi.org/10.1016/j.heares.2013.05.013

Stowe, L. a, & Sabourin, L. (2005). Imaging the processing of a second language: Effects of maturation and prociency on the neural processes involved. *International Review of Applied Linguistics in Language Teaching (IRAL), 43*(4), 329-353. https://doi.org/10.1515/iral.2005.43.4.329

Strauß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed expectancies under degraded speech: revisiting the N400. *Journal of Cognitive Neuroscience, 25(8)*, 1383-95. doi: 10.1162/jocn_a_00389

Stringer, L. M. (2015). Accent intelligibility across native and non-native accent pairings : investigating links with electrophysiological measures of word recognition (Unpublished MPhil dissertation). University College London, London.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, *60*(4),

487–501. https://doi.org/10.1016/j.jml.2009.01.001

Suomi, K., Toivanen, J., & Ylitalo, R. (2008). *Finnish Sound Structure. Phonetics, phonology, phonotactics and prosody.* Oulu: University of Oulu. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Finnish+Sound+Structure.+Phonetics,+phonology,+phonotactics+and+prosody#0

Trubetzkoy, N. (1939). *Principles of Phonology.* (C.A.M. Baltaxe, Berkeley, Trans.). Berkeley, CA: University of California Press.

Tuinman, A., Mitterer, H., & Cutler, A. (2011). Perception of intrusive /r/ in English by native, cross-language and cross-dialect listeners. *The Journal of the Acoustical Society of America*, *130*(3), 1643. https://doi.org/10.1121/1.3619793

Tuomainen, O., Hazan, V., & Romeo, R. (2016). Do talkers produce less dispersed phoneme categories in a clear speaking style? *The Journal of the Acoustical Society of America*, *140*(4), EL320-EL326. https://doi.org/10.1121/1.4964815

Uchanski, R. M. (2008). Clear Speech BT - The Handbook of Speech Perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp.207-235). Oxford: Blackwell Publishing. https://doi.org/10.1111/b.9780631229278.2004.00012.x

Vaden, K. I., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2013). The Cingulo-Opercular Network Provides Word-Recognition Benefit. *Journal of Neuroscience*, *33*(48), 18979–18986. https://doi.org/10.1523/JNEUROSCI.1417-13.2013

Van Alphen, P., & McQueen, J. M. (2001). The time-limited influence of sentential context on function word identification. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 1057–1071. http://dx.doi.org/10.1037/0096-1523.27.5.1057

van Dommelen, W. A., & Hazan, V. (2012). Impact of talker variability on word recognition in non-native listeners. *The Journal of the Acoustical Society of America*, *132*(3), 1690–9. https://doi.org/10.1121/1.4739447

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native- and foreign-accented English: communicative efficiency across conversational dyads with varying language

alignment profiles. *Language and Speech, 53*(4), 510–540. https://doi.org/10.1177/0023830910372495

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America, 121*(1), 519-526. https://doi.org/10.1121/1.2400666

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience, 8*, 1–4. https://doi.org/10.3389/fnhum.2014.00577

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition, 18*(4), 380–393. https://doi.org/10.3758/BF03197127

Walsh, T., & Diller, K., (1979). Neurolinguistic considerations on the optimum age for second language learning. *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society*, 510-524. http://dx.doi.org/10.3765/bls.v5i0.2157

Wagner, M., Shafer, V. L., Martin, B., & Steinschneider, M. (2013). The effect of native-language experience on the sensory-obligatory components, the P1-N1-P2 and the T-complex. *Brain Research, 1522*, 31–37. https://doi.org/10.1016/j.brainres.2013.04.045

Warner, N. (2012). Methods for studying spontaneous speech. In A. Cohn, C. Fougeron, & M. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 621-633). Oxford: Oxford University Press.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167(3917)*, 392-393. https://doi.org/10.1126/science.167.3917.392

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language, 50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Weber-Fox, C. M., & Neville, H. J. (1996). Maturational Constraints on Functional Specializations for Language Processing: ERP and Behavioral Evidence in Bilingual Speakers. *Journal of Cognitive Neuroscience, 8*(3), 231–256. https://doi.org/10.1162/jocn.1996.8.3.231

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceputal reorganization during the first year of life. *Infant Behavior and Development*. 7(1), 49-63. https://doi.org/10.1016/S0163-6383(84)80022-3

White, L., Mattys, S. L., & Wiget, L. (2012). Segmentation cues in conversational speech: Robust semantics and fragile phonotactics. *Frontiers in Psychology, 3*, 1–9. https://doi.org/10.3389/fpsyg.2012.00375

Whitmal III, N. A., Poissant, S. F., Freyman, R. L., & Helfer, K. S. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *The Journal of the Acoustical Society of America*, *122*(4), 2376–2388. https://doi.org/10.1121/1.2773993

Wijngaarden, S. J. Van, Steeneken, H. J. M., Houtgast, T., van Wijngaarden, S. J., Steeneken, H. J. M., & Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native talkers. *The Journal of the Acoustical Society of America*, *112(6)*, 3004–3013. https://doi.org/10.1121/1.1512289

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. *Journal of Neuroscience*, *32*(40), 14010–14021. https://doi.org/10.1523/JNEUROSCI.1528-12.2012

Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csépe, V., … Näätänen, R. (1999). Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. *Cognitive Brain Research*, *7*(3), 357–369. https://doi.org/10.1016/S0926-6410(98)00039-1

Wlotko, E. W., & Federmeier, K. D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia*, *45*(13), 3001–3014. https://doi.org/10.1016/j.neuropsychologia.2007.05.013

Wöstmann, M., Herrmann, B., Maess, B., & Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(14), 3873-3878. https://doi.org/10.1073/pnas.1523357113

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear & Hearing*, *32*(4), 498–510.

https://doi.org/10.1097/AUD.0b013e31820512bb

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., … Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party.” *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Yun, A., Yoon, K., Park, S., Lee, J., Cho, S., Kang, D., … Kim, J. (2015). The Korean Corpus of Spontaneous Speech. *Journal of The Korean Society of Speech Sciences*, *766*(72), 103–109. http://www.dbpia.co.kr/Article/NODE06366731

Appendix 1: Sentence materials (Study 3)

Sentences that were adapted from the Basic English Lexicon (BEL) sentences (Calandruccio & Smiljanić, 2012) to vary final-word predictability

A. Low cloze probability sentences

A lazy worker rests soundly.
My doctor works in that busy neighbourhood.
The state school is large and famous.
The girl loves lemon sweets.
The young performer learned to act.
The couple kissed after fighting.
That shop sells cheap vegetables.
The king and queen planned a funeral.
The foreign tourist was excited and friendly.
The wild horse jumped occasionally.
The strong army won the hill.
The business created many machines.
The father hugs his sad friend.
The wild animals sleep in the cage.
The husband and wife cut the meat.
My grandmother drinks cold milk.
The trees grow sweet oranges.
Their nephew ran around the church.
The English tea smelled strange.
My mother bakes delicious pasta.
The fish swam slowly in the water.
The chef prepares breakfast in the hotel.
The warm sunshine felt fantastic.
The restaurant sells red cherries.
The fat pig slept on the carpet.

She drove the bus down the mountain.

The black cat climbed the wall.

The sad pets need friends.

A kind word is always good.

The fried egg was cooked in seconds.

A lazy child sleeps frequently.

The grape juice spilled on the dress.

The best explanation is often obvious.

Their famous son danced secretly.

The best worker went on the tour.

That tiny animal is cute but filthy.

The crowd watched the talented woman.

The spoiled potatoes tasted terrible.

The lonely duck swims in the pool.

The cool night was comfortable and calm.

They played fast music on the balcony.

The bedroom rug had a large border.

The map shows the main buildings.

The old rubbish attracts animals.

The flags fly high and grand.

The twin sisters watched a fly.

The driver stopped suddenly in the rain.

The long project was completed on budget.

Those little kids are tired again.

He screamed loudly in the crowded restaurant.

The cherry pie was warm and fresh.

The excited children cheered for their uncle.

The waiter broke ten bottles.

The ocean looked perfectly peaceful.

The divorced couple sat at the bar.

The group heard slow drips.

The kind lady gives oranges.

The coffee cake was a perfect gift.
That new student is quiet and clever.
The twins received the same letter.
A hungry rabbit eats everything.
The first question was confusing and long.
My grandfather made wooden plates.
The friendly baby hugs her.
The artist studies Italian and Russian.
The helpful nanny cleaned the car.
The teacher chooses difficult books.
The soft music pleased them.
The shy guest speaks English.
The scared mouse stayed in the garden.

B. Anomalous sentences

The weak plant is barely opposed.
The red vegetables grow in the boyfriend.
The plane will land in ten windows.
Our teacher answers every peanut.
The cricket ball flew across the noise.
A foreign country is exciting to marry.
His parents tell boring bananas.
The proud fans cheered for their sea.
His girlfriend loves Chinese sleep.
The thirsty cat drank nails.
The three cousins did their math pancake.
The white horse lives on a finger.
The metal key opened the news.
The couple lives a peaceful bang.
The mouse found tasty hug.
The vegetables grew in the green payback.

The only hotel is far and forensic.

The sick neighbour asks for jail.

His speech was boring and too red.

He cut the steak with a dog.